Drawing impossible boundaries: field delineation of Social Network Science

Lietz, Haiko

Empfohlene Zitierung / Suggested Citation:
Lietz, H. (2020). Drawing impossible boundaries: field delineation of Social Network Science. *Scientometrics*, *125*(3), 2841-2876. https://doi.org/10.1007/s11192-020-03527-0

Scientometrics (2020) 125:2841–2876
https://doi.org/10.1007/s11192-020-03527-0

Drawing impossible boundaries: field delineation of Social Network Science

Haiko Lietz[1]

Received: 20 November 2019 / Published online: 13 June 2020

Abstract
"Big" digital behavioral data increasingly allows large-scale and high-resolution analyses of the behavior and performance of persons or aggregated identities in whole fields. Often the desired system of study is only a subset of a larger database. The task of drawing a field boundary is complicated because socio-cultural systems are highly overlapping. Here, I propose a sociologically enhanced information retrieval method to delineate fields that is based on the reproductive mechanism of fields, able to account for field heterogeneity, and generally applicable also outside scientometric, e.g., in social media, contexts. The method is demonstrated in a delineation of the multidisciplinary and very heterogeneous Social Network Science field using the Web of Science database. The field consists of 25,760 publications and has a historical dimension (1916–2012). This set has high face validity and exhibits expected statistical properties like systemic growth and power law size distributions. Data is clean and disambiguated. The dataset with 45,580 author names and 23,026 linguistic concepts is publically available and supposed to enable high-quality analyses of an evolving complex socio-cultural system.

Keywords Field delineation · Sociologically enhanced information retrieval · Boundary problem · Social Network Science (SNS) · Web of Science

Introduction

High-quality research rests on high-quality datasets. "Big" digital behavioral data consists of traces of behavior left by uses of, or harnessed by, digital technology. It is often created for economic purposes and increasingly allows large-scale and high-resolution network analyses of the behavior and performance of persons or aggregated identities in whole fields (Lazer and Radford 2017). A field or network domain is comprised of a story set

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11192-020-03527-0) contains supplementary material, which is available to authorized users.

✉ Haiko Lietz
haiko.lietz@gesis.org

[1] GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6–8, 50667 Cologne, Germany

(domain) and the persons collectively enacting it (network). Put simply, a field is a set of persons thematically concerned with a set of things (White 2008). *Field delineation* is the computational task of collecting, or retrieving from a database, the building blocks of fields (Zitt 2015). Depending on how a field is represented by data, those blocks can be as diverse as publications or tweets. But there is a conceptual problem: the goal of field delineation is to draw a boundary that does not exist in reality. Socio-cultural systems have no clear-cut boundaries but are highly overlapping (Palla et al. 2007) due to their constructed and fractal nature (Abbott 2001; Fuchs 2001). This is known as the *boundary problem* in sociology. In that sense, to delineate a field is to draw an impossible boundary.

Bibliographic data is not only a very early example of unobtrusive behavioral data as publications are not produced for the purpose of statistical analysis. It is also a form of multiplex data (Padgett and Powell 2012) for joint analyses of network (co-authorship) and domain (citation and word usage). This makes it sociologically very appealing. Bibliographic databases like the Web of Science and Scopus provide classification systems to aid publication retrieval. Since these systems classify journals, not publications, along coarse disciplinary lines, they are of limited help when it comes to delineating interdisciplinary fields that span across these journal classes. Article-level classification systems can improve the fine-grained publication retrieval of interdisciplinary fields (Glänzel and Schubert 2003; Neuhaus and Daniel 2009; Waltman and van Eck 2012; Sjögårde and Ahlgren 2018). However, they may not be available other than to institutes with priviledged data access (Waltman and van Eck 2012) or may be of limited trustworthiness due to their black box nature (Sinha et al. 2015). Furthermore, even if an article-level classification system is available and trustworthy, publication retrieval still involves manual checking and refinement steps (Milanez et al. 2016).

I propose a *sociologically enhanced information retrieval* method for field delineation with three parameters that is tailored to retrieving substructured fields, does not rely on an existing classification system, is rooted in sociological theory, and can be applied in non-scientometric settings. For example, it is supposed to be capable of retrieving publications from a bibliographic database as well as tweets from a Twitter corpus. My method is based on the bibliometrically enhanced information retrieval method of Zitt and Bassecoulard (2006) according to which a field is delineated by starting with a precise seed set of publications, then identifying its core cited references, and finally retrieving publications that cite this core. This citing/cited/citing logic is a good starting point because it resonates with the mechanism how complex socio-cultural systems operate via feedback (White 2008; Padgett and Powell 2012). On the way to a general field delineation method, the citation-based retrieval method is generalized to include word usage (Zitt 2015) and subfield delineation is introduced to deal with field heterogeneity (Mogoutov and Kahane 2007). The solution to the boundary problem is to classify a sample set of transactions (e.g., publications or tweets), decide how many false positives or false negatives one is willing to accept in retrieval, and specify the respective fuzziness of the boundary.

This procedure is demonstrated in a delineation of the Social Network Science (SNS) field using the Web of Science database. This field is defined as the network domain that studies socio-cultural systems in a relational way—a multidisciplinary science of social networks, not a sociological network science. As such it roughly combines the classical Social Network Analysis (SNA) field (Freeman 2004) and the subfield of Network Science (Barabási 2016) that studies socio-cultural systems. SNS is a particularly interesting case because it is an evolving field that has seen many twists and turns (Hummon and Carley 1993; Freeman 2004; Garfield 2004; Shibata et al. 2007; Leydesdorff et al. 2008; Lazer et al. 2009; Brandes and Pich 2011; Freeman 2011; Lancichinetti and Fortunato

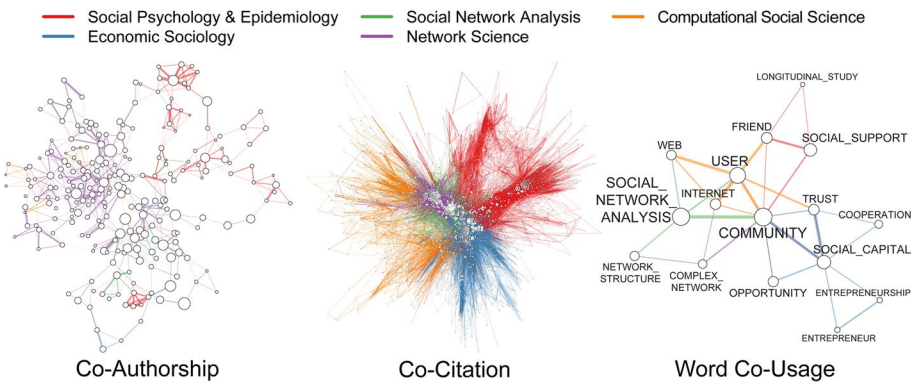


Fig. 1 Co-selection graph cores of Social Network Science. From left to right: authors co-authoring, references co-cited in, and words co-used in publications. Colors indicate subfields and unveil how the latter overlap in the co-selection of facts. See description of the final dataset for how these graphs are constructed. (Color figure online)

2012; Batagelj and Cerinšek 2013; Hidalgo 2016; Maltseva and Batagelj 2019). So far, the only bibliographic dataset of the whole field is the SN17 dataset retrieved from the Web of Science by Maltseva and Batagelj (2019). It is based on the SN5 dataset (Batagelj and Cerinšek 2013) retrieved from the Web of Science in 2007. SN5 contains publications that use the search term `SOCIAL NETWORK*` in either title, abstract, or keywords, or have been published in the journal *Social Networks*, plus the "most frequently cited works" of those publications.[1] SN17 is an extension of SN5 to the year 2018 using the same search term but adding new complete network-related journals (Maltseva and Batagelj 2019).

The goal of the delineation task is to create a high-quality dataset that has undergone manual oversight. It should exclude publications that talk of "social networks" metaphorically, have disambiguated author names, contain the most important citations made in publications' reference lists (not just to items in the database), include multi-token linguistic concepts (*n*-grams), and allow historical analysis, i.e., capture the field from its predecessors on. The SN17 dataset does not meet these criteria. Its boundary is too fuzzy because it includes publications that use the networks term metaphorically. Therefore, I have delineated SNS anew. The resulting dataset consists of 25,760 biographical records retrieved from the Web of Science, ranging from 1916 to 2012. There are 45,580 distinct authors, 574,036 cited references, and 23,026 linguistic concepts. Except for citations, the data is made available to the community under a Creative Commons license (Lietz 2019) and can be explored online in a virtual Jupyter Notebook without the need to install or master a programming language (Lietz 2020). Figure 1 gives an impression of the networks that can be constructed from this dataset.

This paper is a revised chapter of my dissertation (Lietz 2016). In the next section, the sociological model of fields is introduced. Then I describe the field delineation procedure in detail before I apply it to delineating the SNS field. A discussion and conclusion is offered

[1] The SN5 dataset can be downloaded at http://vlado.fmf.uni-lj.si/pub/networks/pajek/WoS2Pajek/WoS2Pajek.htm, visited September 18th, 2019.

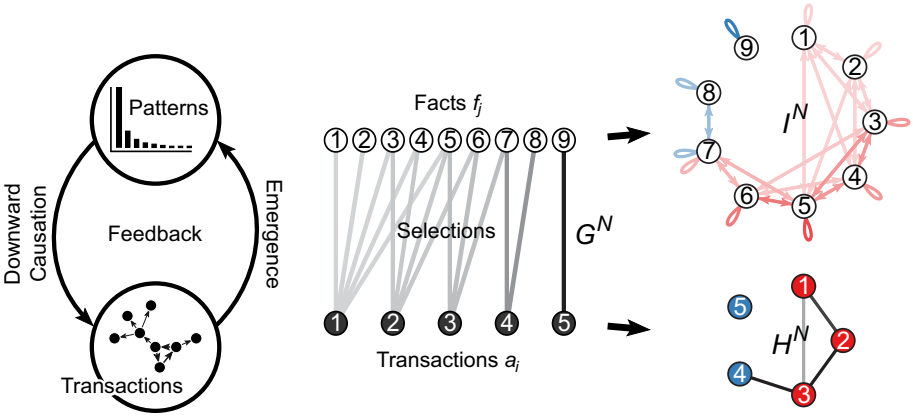


Fig. 2 Unified field and data model. The feedback mechanism of field reproduction is depicted on the left. Transactions are the constituents of fields and facts are the components of the emergent patterns that influence transactions in downward causation. This field model maps to a bipartite graph model of selections, shown in the middle. Selection matrices G can be projected into fact-coupled transaction matrices H (which are used for partitioning seeds and fields) and into fact co-selection matrices I (as shown in Fig. 1), both depicted on the right. For the citation practice, transactions are publications, facts are references, selections are citations, H is the reference-coupled "bibliographic coupling" publication matrix, and I is the reference co-citation matrix which represents the conceptual pattern of the field. The graph plots visualize a toy example where normalization is used throughout. Consult the "Technical Appendix" for details of the formalism

in the last section. Most of the mathematical formalism is put in a "Technical Appendix". Supplementary Information is given for data processing and publication classification.

Sociological field model

The field delineation procedure is supposed to generate data that resembles the operations of persons in the network domain to be delineated. Therefore, it is necessarily rooted in a behavioral model. In sociology, the *field* concept refers to a structure of positions that are equipped with different sorts of social capital (Bourdieu and Wacquant 1992). This concept is compatible with the concept of *network domain* in Relational Sociology (Schmitt 2019). Throughout the paper, I use these terms interchangeably. There is never a social network without a culture giving meaning to connectivity, and, vice versa, there is never a culture without it being practiced in social relations. The concept of network domain captures this duality of connectivity (network) and culture (domain) (White 2008). For this reason, I refer to "socio-cultural systems" as opposed to the more common "social systems" term.

Network domains reproduce themselves in self-organization. *Transactions* are their building blocks (Emirbayer 1997). In the "network" dimension, these are social relations. In the "domain" dimension, facts are selected. Durkheim (1982 [1895]) conceptualized a *fact* as a thing that emerges from collective action and influences individual behavior. *Selection* expresses this duality that persons both actively chose to make reference to ("select") facts and, at the same time, are influenced by them. Put into the relational perspective of complex socio-cultural systems, a field operates by persons making selections in transactions from which meaning structures emerge which feed back onto future

transactions (Breiger 1974; Fuhse 2009; Padgett and Powell 2012; Page 2015). The feedback loop of field reproduction is depicted in the left part of Fig. 2. While emergence is non-causal, "downward causation" conceptualizes the causal part of the feedback dynamic (Flack 2017). *Meaning structures* are any kind of observable pattern, like fact co-selection structures or fact size distributions. They have the function to signal which fact belongs to the core of the network domain. The core harbors the agreed-upon concepts and institutions of a network domain (Fuchs 2001). Facts can be distinguished according to the capability of *agency*, to actively engage into social action (Emirbayer and Mische 1998). If facts are capable of agency, e.g., persons, groups, or organizations, the corresponding meaning structure is social. Meaning structures built of symbols, words, ideas etc. are cultural networks (McLean 2017). Finally, network domains involve multiple *practices* or types of agency (Swidler 1986).

Sociologically, bibliographic data is particularly interesting because it contains data on three practices: one social and two cultural. *Authorship* is the social practice of communicating research results in scholarly publications—in my terminology: authors are selected in publications. The other practices are cultural because the facts are not capable of agency. *Citation* is the practice of making reference to concept symbols, i.e., references are cited in transactions; *word usage* is the practice of language, i.e., words are selected in transactions.

Core concepts in scientometrics are easily incorporated into this field model. For example, the duality of connectivity and culture is mirrored in the idea that research communities are not just social groups but "thought collectives" (Fleck 1979 [1935]) who "share similar research interests" (Zuccala 2006, p. 155). A publication is a transaction made by authors in which references and word concepts are selected (cited and used). A cited publication is a fact since, being a concept symbol, it influences the citing publication (Small 1978). Co-citation (Small 1973) and co-word (Callon et al. 1986) networks are examples of cultural meaning structures for the practices of citation and word usage, respectively. The size distributions of Lotka (1926), Bradford (1985 [1934]), Zipf (2012 [1949]), and Price (1976) are descriptions of such meaning structures, signaling who are the core scholars, journals, linguistic concepts, and citeable references, respectively, in a field.

Field delineation procedure

The procedure proposed here is based on the bibliometrically enhanced publication retrieval procedure of Zitt and Bassecoulard (2006). This method can be mapped to the field model just described. It is based on the citing/cited/citing logic that publications which are known to belong to a field of interest cite a set of core references which must also be cited by other field publications. One starts on the citing side of transactions: a field is delineated by retrieving, from the set of all publications S in a database, a seed set A, using expert-defined lexical queries that are very precise. Then one moves to the cited side of meaning structures: from A, the set B of cited references is identified in which a reference is cited y times; to obtain a generic and specific core of cited references, B is reduced to C by requiring that references in B receive $y \geq Y$ citations from the seed publications; Y is a genericness parameter; next C is reduced to the "cited core" D by requiring that references in C receive a fraction $u = y/y' \geq U$ of their citations from the seed A; y' is the number of citations a reference receives in the whole database S; U is a specificity parameter. Finally, one goes back to the citing side: the field E is the set of publications that each cite

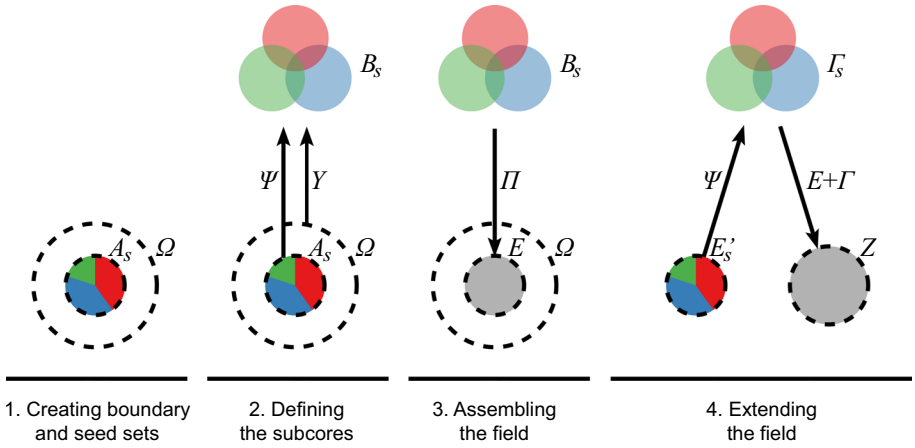


Fig. 3 Field delineation procedure. First, a set of candidate transactions (boundary set) Ω and an initial iteration of the field (seed set) A are created, and A is partitioned into subfields A_s. Second, the overlapping subcores B_s selected by A_s are defined using threshold parameters Ψ for fact genericness and Y for fact specificity. Third, the field E is assembled by retrieving those transactions from Ω that select the subcores B_s and fulfil a stated requirement (minimum subcore recall of precision), defined by the assembly parameter Π. Fourth, the final (extended) field set Z is created by adding to E its cores Γ_s if transactions and facts are of the same entity type. Consult Table 1 for details on the notation used in the procedure

at least $x \geq X$ references in the cited core D; X is a relevance parameter. Throughout the paper, I refer to this procedure as the *original method*.

I generalize this procedure to be able to delineate any type of field defined as the feedback process of transactions selecting facts. The reasons to also modify the original method are twofold. First, it is unfair in the case of field heterogeneity. For a subfield that is large or has a very skewed citation distribution, $Y = 10$ citations may not be much, but for a subfield that is small or has a less skewed size distribution, it may be a lot. Even for subfields with similar size and skewness, thresholding on a particular Y would be unfair if reference list lengths vary. To mitigate this problem, I introduce a clustering sub-procedure and perform field delineation on the subfield level. Second, having access to a whole database of all transactions S is the exception rather than the rule (e.g., in the case of commercial databases like the Web of Science or Scopus), if not impossible. Often, database access is restricted or download limits are imposed. My method does not require access to a full database. Instead, the field is built from a restricted set of candidate transactions. As a consequence of that modification, expert knowledge is not needed in the first delineation step of creating the seed but in a later step, and the risk of expert bias is minimized. The following modified field delineation procedure is sketched in Fig. 3. The notation used throughout this paper is summed up in Table 1.

Creating boundary and seed sets

The first step is to create two sets of transactions. The *boundary* set Ω contains the transactions that are candidates for belonging to the field (being inside the boundary). It should be devoid of transactions that are completely off topic because, in the next step, a sample will be coded as inside/outside the boundary, and this classification task should decide upon nuances, not obviousness. The *seed* set A is the first iteration of the field. It should be as precise as possible as it is used to create candidate lists of core facts. But it needs not be as precise as the seed in the original method because expert knowledge is involved in the

Table 1 Notation of field delineation procedure

Entities	
a_i	Transaction: building block of fields
f_j	Fact: thing selected in, or descriptor of, transactions
Fact properties	
$\psi_{j,s}$	Genericness of fact j in subseed s; cumulative sum of ranked selection fractions; obtained by ranking facts j descendingly by the fraction of transactions in s that select them and taking the cumulative sum
$\upsilon_{j,s}$	Specificity of fact j in subseed s; fraction of transactions fact j retrieves from Ω' that are relevant
Transaction sets	
Ω	Boundary: transactions that are candidates for being in E
Ω'	Sample of Ω where transactions must be coded as inside/outside the boundary (relevant/irrelevant for the field)
Ω'_{in}	Subset of Ω' coded as inside the boundary (relevant)
A	Seed: first iteration of the field
A_{s}	Subseeds: first iteration of the subfields; subsets of A; obtained by detecting non-overlapping communities in H^{N}_A
E_{s}	Second iteration of the subfields; obtained by retrieving from Ω all transactions that select a fact in B_{s}
E	Second iteration of the field; set union of E_{s}
E'_{s}	Subsets of E; obtained by detecting non-overlapping communities in H^{N}_E
Z	Third and final iteration of the field; set union of E and Γ_{s}
Fact sets	
B_{s}	Subcores: set of facts j selected in subseeds A_{s}; defined by $\psi_{j,s} \leq \Psi$ and $\upsilon_{j,s} \geq \Upsilon$
Γ_{s}	Set of facts j selected in subsets E'_{s}; defined by $\psi_{j,s} \leq \Psi$
Parameters	
Ψ	Genericness parameter; sets efficiency of retrieval
Υ	Specificity parameter; sets accuracy of retrieval
Π	Assembly parameter; must be chosen to be minimum recall or minimum precision of subcores B_{s}
Matrices	
G	Selection matrix: Bipartite matrix of transactions i and facts j
G^{N}	Normalized selection matrix: Row-normalized matrix G
H	Fact-coupled transaction matrix: First projection of G
H^{N}	Normalized fact-coupled transaction matrix: First projection of G^{N}
I	Fact co-selection matrix: Second projection of G
I^{N}	Normalized fact co-selection matrix: second projection of G^{N}

Compare with the field delineation procedure schematized in Fig. 3

classification just mentioned. Ω is a superset of A, i.e., the seed is fully contained in the boundary set.

To account for field heterogeneity—the existence of differently sized subfields or of varying selection practices—, two actions are taken. First, the weights of the selections made in a transaction are normalized to sum to unity (Batagelj and Cerinšek 2013). To handle data, a unified field and data model is introduced which maps the field model of transactions and facts to a bipartite graph model of selections (Fig. 2). In a nutshell, bipartite

selection graphs consist of two types of vertices with only inter-type connections. The first type of vertices are transactions; the second type of vertices are facts; an edge is created if a fact is selected in a transaction. For each practice, one normalized selection graph is constructed. Details are laid out in the "Technical Appendix".

Second, delineation is made on the subfield level, i.e., subseeds A_s are identified. This action is inspired by Mogoutov and Kahane (2007). The goal is to create clusters of transactions based on similar selection profiles (Doreian et al. 2004). There are many ways to construct such similarities (Eck and Waltman 2009). Here, a purely graph-theoretic approach is used that has very natural interpretations. It results in analytical transaction graphs where edge weights resemble transaction similarities in the [0, 1] interval (cf. "Technical Appendix"). Given this graph, the seed is partitioned (single membership) using community detection (Fortunato 2010). Refined computational methods should proceed by detecting dynamic communities.

In the case of scientometrics, when publications are coupled through the citation practice, this is the "bibliographic coupling" network (Kessler 1963). Other transaction graphs are possible, e.g., author-coupled and word-coupled publication graphs. The way selections are normalized had first been proposed by Leydesdorff and Opthof (2010) for counting citations.

Defining the subcores The second step is to identify a subcore B_s for each subseed A_s. Subcores must be both generic and specific. In the original method, *genericness* is ensured through requiring core facts to each have at least a certain number of selections from the seed. If one chose the same absolute threshold for all subfields, then small subfields and those with a less skewed size distribution would be punished. To ensure that all subfield cores are equally generic, my method takes advantage of the situation that few facts are selected by, or retrieve, many transactions.

Put shortly, facts $f_{j,s}$ are ranked such that a fact's rank increases when it is highly selected in a subseed s but decreases when it is highly selected in the whole seed ($tf * idf$ principle). The genericness $\psi_{j,s}$ of a fact j in subseed s is then the cumulative sum of selection fractions K^{N}.[2] Finally, facts are thresholded against a genericness parameter Ψ, the first parameter, such that $\psi_{j,s} \leq \Psi$. For example, when $\Psi = 0.1$, then the highest-ranked facts that accumulate no more than the top ten percent of all selections are chosen to constitute a subcore. The exact method is laid out in the "Technical Appendix".

To ensure *specificity*, informed manual work must be involved in the delineation procedure at some point. Zitt and Bassecoulard (2006) have proposed to define those facts as belonging to the core that receive at least a certain fraction of their selections from a seed that is highly precise. This requires expert knowledge in the first step of defining the seed. This knowledge can result in a lexical query that does not, or hardly, retrieves false positives or in the identification of curated collections, e.g., conference proceedings or tweet collections, where all transactions are on topic.

Here, I propose an alternative approach: to attribute a specificity to facts, a sample Ω' of the boundary set Ω is coded along an inside/outside dichotomy. This approach changes the kind of expert work from defining a transaction set to defining a codebook on how to classify transactions. Having this codebook, the actual classification task can be outsourced—maybe even to crowd workers if they are well trained and paid. The specificity of

[2] In words, the genericness of the first-ranked fact amounts to its selection fraction; the genericness of the second-ranked facts amounts to the sum of the selection fractions of the two highest-ranked facts; the last-ranked fact has unit genericness.

fact $f_{j,s}$ is then $v_{j,s} = |\Omega'_{\text{in}}|_{j,s}/|\Omega'|_{j,s}$. Here, $|\Omega'_{\text{in}}|_{j,s}$ is the size of the subset of transactions in the sample, retrieved by $f_{j,s}$, that are ruled to belong to the field, and $|\Omega'|_{j,s}$ is the size of the subset of transactions in the sample retrieved by $f_{j,s}$. Finally, facts are thresholded against a specificity parameter Y, the second parameter, such that $v_{j,s} \geq Y$. For example, when $Y = 0.5$, then a fact is chosen to co-constitute a subcore if at least half of the transactions it retrieves are relevant for the field (ruled inside the boundary).

Genericness ensures that retrieval is efficient. The larger this parameter is set, the more a subcore consists of large numbers of less selected facts. Specificity ensures that retrieval is accurate. The larger this parameter is set, the more a subcore consists of facts that only retrieve relevant transactions as judged by the coding of the sample Ω'. Subfield retrieval is evaluated using recall and precision. Recall $= |\Omega'_{\text{in}}|_{D_s}/|\Omega'_{\text{in}}|$ is the fraction of relevant transactions in the sample Ω' that are retrieved by the subcore D_s. Precision $= |\Omega'_{\text{in}}|_{D_s}/|\Omega'|_{D_s}$ is the fraction of transactions retrieved by D_s from the sample Ω' that are relevant. The evaluation metrics and retrieval parameters are naturally related. Recall is a transaction-side measure and is strongly influenced by the fact-side genericness parameter; precision is a transaction-side measure and is strongly influenced by the fact-side specificity parameter.

Assembling the field

The third step is to create the second iteration of the field by retrieving those transaction sets E_s that select the subcores defined in the previous step and creating the set union E. One can use a different retrieval parameter setting for each subfield. Then the task is to decide on a setting by trading recall off against precision—how many false positives is one willing to accept for the benefit of reducing false negatives? If the goal is to delineate a field through one set of facts chosen by one parameter setting, not one set and setting for each subfield, then the problem arises that a particular parameter setting can entail varying recall and precision for different subcores. This is because a fact that belongs to the core of one subfield can belong to the periphery of another subfield. Then, define an assembly parameter Π, the third parameter: a minimum recall or precision that applies to all subfields alike. From this minimum value, a universal parameter setting can be deduced that maximizes overall precision or recall.

Extending the field

The fourth and final step is originally not intended and only makes sense if transactions and facts are of the same entity type. For example, publications and cited references are of the same type but publications and used words are not; tweets and retweeted tweets are of the same type but tweets and used hashtags are not. The step consists of partitioning E into subfields E'_s using community detection as in the first step, defining the subcores Γ_s selected by E'_s, and adding to E the facts in Γ_s, thresholding on the value of Ψ identified in the previous step. Call this third and final iteration of the field the extended field Z. In the scientometric case, this step resembles adding to the field its most cited references because those have important meanings even though they may not be directly related to the topic. This is often the case for methodological contributions.

Delineating Social Network Science

As stated in the introduction, the goal is to delineate SNS as a multidisciplinary science of social networks that roughly combines the classical SNA field and the subfield of Network Science that studies socio-cultural systems. Data was queried from the Web of Science. I chose this database because its records are historical (they go back to 1900), they

are systematically collected via journals with not low impact factors (Garfield 1979), and because a lot of effort is put into upholding a high data quality. The Microsoft Academic Graph is also historical and it automatically collects many more records (Sinha et al. 2015), but for that reason its data quality is also lower. Queries were made in 2013 via the online interface at www.webofknowledge.com. Unfortunately, records can only be downloaded in batches of 500. This complicates field delineation enormously and has caused me to delineate SNS on the subfield level but not dynamically.

Creating boundary and seed sets

In this first step, the boundary set Ω, from which publications representing SNS are "recruited", and the seed set A, the first iteration of the field that selects the subcores later used for assembling the field, are created. On the one hand, candidate publications should not be required to use the word `SOCIAL NETWORK*` in title, abstract, or author keywords (throughout the paper, "words" are meant to include sequences of n tokens or n-grams) because a contributions to SNS may well use a different word (e.g., "social relation"). On the other hand, not all publications using the `SOCIAL NETWORK*` 2-gram should automatically be inside the boundary. For example, "social network" is also used metaphorically, in the case of which I do not consider the respective publication to be inside SNS. But all candidates for E, the second iteration of the field created in the third step, should use the words `SOCIAL` and `NETWORK*`. These thoughts define the two initial sets. The boundary set Ω contains 44,308 publications using the words `SOCIAL` and `NETWORK*`. The seed A is a subset of Ω and contains 23,568 publications using `SOCIAL NETWORK*`. Note that the seed is not very precise. Publication years in Ω range from 1953 to 2014.[3]

This data was then processed. Each publication and reference was transformed into a key such that a cited reference can be matched to a citing publication. Granovetter's (1973) paper, e.g., has the matchkey `GRANOVET_1973_A_1360`. All titles, abstracts, and author keywords were preprocessed and stemmed. All words used by at least one author in the seed as a keyword represent the vocabulary. A vocabulary word is selected by a publication if it is used in either the title, abstract, or author keywords. For details of data processing see the Supplementary Information (Section 1).

Based on the description of Scott (2012) and other analyses of SNS (Hummon and Carley 1993; Freeman 2004; Shibata et al. 2007; Lazer et al. 2009; Brandes and Pich 2011; Freeman 2011; Hidalgo 2016; Maltseva and Batagelj 2019), the field is expected to have a social-psychological path with a strong graph-theoretical focus, a diverging ethnographical lineage, a structuralist narrative following the breakthroughs of White et al. in the 70s, a development driven by physics starting around 2000, and a recent surge of research on animal social networks. These paths belong to different scientific disciplines with different styles of practice. Therefore, SNS is not delineated as if it had

[3] Sets were retrieved on November 6th, 2013. Publications have been delivered for the Science Citation Index Expanded (publication years since 1900), Social Sciences Citation Index (since 1900), Arts and Humanities Citation Index (since 1975), Conference Proceedings Citation Index—Science (since 1990), and Conference Proceedings Citation Index—Social Science and Humanities (since 1990). The query for the boundary set is `TS=(SOCIAL and (NETWORK or NETWORKS))` and that for the seed is `TS=(SOCIAL NETWORK or SOCIAL NETWORKS)`. That means, publications on "social networking" are initially excluded. Publications where the search terms only occur as *KeyWords Plus* have been filtered out ex post because these keywords were found to be unreliable. Results include all document types (articles, reviews, letters, editorials, corrections, etc.).

Table 2 Seed clustering statistics for different coupling methods

	Network		Clustering	
	Density	LCC (%)	Modularity	Consensus
A	0.0002	41.1	0.96	0.78 ± 0.03
R	0.0332	93.0	0.46	0.95 ± 0.02
AR	0.0333	96.1	0.60	0.92 ± 0.02
W	0.5315	99.6	0.14	0.97 ± 0.01
AW	0.5315	99.7	0.14	0.79 ± 0.18
RW	0.5414	99.8	0.14	0.92 ± 0.14
ARW	0.5414	99.9	0.15	0.81 ± 0.17

Rows list methods of coupling publications through authors (A), cited references (R), used words (W), and combinations thereof. In the case of combinations, edge weights are averaged. Network density increases from top to bottom, the author-coupled publication graph is three orders of magnitude sparser than when publications are coupled through word co-usage. In all networks that involve word coupling, every second edge is actually realized (density). Accordingly, the size of the largest connected component (LCC) is only 41.1% for author coupling but 99.6% for word coupling. The internalization of edges to subseeds (Modularity Q) decreases with increasing density. Consensus (means and standard deviations) tells how much partitions from clustering are reproducible. We use the Adjusted Rand Score, a standard measure for the similarity of two partitions, to compute consensus. It counts similarly partitioned pairs of publications and compares the result to a null model. 1 (− 1) means that two partitions are maximally identical (unidentical)

one core but by accounting for the heterogeneity of subfields with possibly different sizes and publication characteristics.

To partition the seed, I created a selection graph each for the three practices of authorship, citation, and word usage. For the latter, I did not distinguish whether a word is used in title, abstract, or author keywords. *KeyWords Plus* generated automatically from reviewing reference titles (Garfield and Sher 1993) are not used in this study. The three selection graphs were then projected into fact-coupled transaction graphs following the method depicted in Fig. 2 and described in the "Technical Appendix". These graphs and their combinations were then clustered using Louvain community detection (Blondel et al. 2008).

Table 2 shows that graphs from author coupling are two orders of magnitude more sparse than from reference coupling and three orders more than from word coupling. As a consequence, they also differ largely in how many publications belong to the largest connected component (LCC). For author coupling, only 41.1% of all publications are at least indirectly similar via shared authors. The three types of facts also have a different power of distinction. Modularity Q quantifies the extent to which edges are internalized to clusters (Newman 2006), i.e., how permeable subseed boundaries are. A modularity of $Q_{\text{ref}} = 0.46$ for reference coupling means that cited references are less distinctive than authors ($Q_{\text{aut}} = 0.96$) but more than words ($Q_{\text{wrd}} = 0.14$). Rows for hybrid coupling indicate that, once words are part of the coupling mix, Q is low, i.e., subseeds are largely overlapping. This is nothing else but the fact that words obtain their meaning in co-usage, language can be flexibly used, and is less precise in delineating fields than citation (Glänzel and Thijs 2011; Zitt 2015).

Reproducibility is another issue. Louvain community detection has a stochastic element (Lancichinetti and Fortunato 2012). Intuitively, the more boundaries are overlapping, the more publications will be assigned to a partition based on chance. The Adjusted Rand Score quantifies how similar two partitions are (Fortunato 2010).[4] I arrive at means and standard deviations by comparing the solutions of ten runs. It turns out—counter-intuitively—that clustering word-coupled graphs is most and author-coupled graphs is least reproducible. This is because there is less randomness in partitioning lexical graphs as similarity scores (edge weights) have a much wider spectrum.

Summing up until here, even though the component communities of word-coupled publication graphs are most strongly overlapping, their partitions are most reproducible. But once hybrid coupling is used, including references, authors, or both, reproducibility drops. It is clear that all further results are contingent on the choice of facts for coupling publications. At this point, I decided to exclude author coupling from the following considerations because it decreases reproducibility. But there is also a substantive argument: the cultural and the social operate on different time scales. Words, references, and their co-selections are much more institutionalized than authors and team compositions (Padgett and Powell 2012). From this perspective, not coupling publications via authors means not allowing social currents to have an impact on subsequent results and aiming at a more culture-dependent analysis.

In Table 3, the communities or subseeds from reference, word (lexical), and reference/word (hybrid) coupling are described via rankings like top subject categories and facts for the corresponding selection graphs. From the ten clustering runs, the one with the largest modularity is used. Partitions are robust in that they—with one exception—describe five non-trivial communities. By interpreting subseed descriptions, I label these Social Psychology and Epidemiology (SPE), Economic Sociology (ES), Social Network Analysis (SNA), Network Science (NS), and Computational Social Science (CSS). Partitioning for different fact coupling also results in the same temporal ordering. SPE is the oldest community and CSS is the newest. The choice of coupling has an effect on subseed composition. SPE is much larger when delineated lexically or the hybrid way; ES is smaller. Reference coupling results in two subseeds for SNA.

Since no gold standard exists, there is no objective criterion to evaluate the partitions. I chose hybrid coupling because hybrid methods balance the advantages and disadvantages of citation-based and lexical approaches (Braam et al. 1991; Glänzel and Thijs 2011; Zitt 2015). Having excluded author coupling, this prevents either references or words from determining future results.

Defining the subcores

In this second step, the subcores B_s are defined from which the field is later assembled. To obtain the subcores, the genericness and specificity of each fact (reference and word) is determined for each subseed. Fact genericness $\psi_{j,s}$ can directly be computed from the normalized selection matrices of the subseeds (cf. "Technical Appendix"). To obtain fact specificity, I took a sample Ω' of 1000 publications from the boundary set, 499 of which are in the seed, and manually decided if they should be inside or outside SNS

[4] The Adjusted Rand Score is biased in the case of unequal, unbalanced partitions. The Normalized Mutual Information score also discussed by Fortunato (2010) does not have this disadvantage.

Table 3 Description of subseeds from different coupling methods

	Reference	Word	Hybrid
(a) *Social Psychology and Epidemiology*			
Publications	5088	6885	6850
Year quartiles	1999/2007/2011	2001/2008/2011	2001/2008/2011
Subject category	Classifications		
Public, env. and occup. health	1041 (1)	1197 (1)	1216 (1)
Psychiatry	732 (2)	753 (2)	761 (2)
Gerontology	433 (3)	536 (3)	539 (3)
Psychology, multidisciplinary	380 (4)	440 (4)	445 (4)
Social science, biomedical			361 (5)
Psychology, developmental		345 (5)	
Psychology, clinical	304 (5)		
Author keyword	Usages		
SOCIAL_SUPPORT	527 (1)	561 (1)	567 (1)
DEPRESSION	143 (2)	145 (2)	145 (2)
HIV	117 (3)	118 (5)	121 (5)
SOCIAL_CAPITAL		190 (3)	177 (3)
GENDER		129 (4)	131 (4)
QUALITY_OF_LIFE	99 (4)		
AGE	95 (5)		
Reference	Citations		
BERKMAN_1979_A_186	334 (1)	322 (1)	331 (1)
COHEN_1985_P_310	252 (2)	257 (2)	261 (2)
RADLOFF_1977_A_385	234 (3)	230 (3)	233 (3)
HOUSE_1988_S_540	219 (4)	224 (4)	227 (4)
GRANOVET_1973_A_1360		390 (5)	346 (5)
COBB_1976_P_300	173 (5)		
Author	Authorships		
LATKIN,_CA	51 (1)	47 (1)	49 (1)
CHRISTAKIS,_NA	31 (4)	43 (2)	40 (2)
LITWIN,_H	32 (2)	34 (3)	34 (3)
LATKIN,_C	28 (3)	27 (4)	28 (4)
BERKMAN,_LF	26 (5)	25 (5)	25 (5)
(b) *Economic Sociology*			
Publications	4983	3963	3780
Year quartiles	2006/2009/2011	2005/2009/2011	2005/2009/2011
Subject category	Classifications		
Management	789 (2)	590 (1)	598 (1)
Sociology	856 (1)	511 (2)	514 (2)
Business	538 (3)	426 (3)	430 (3)
Economics	448 (4)	263 (5)	263 (5)
Geography	284 (5)	237 (4)	245 (4)
Author keyword	Usages		
SOCIAL_CAPITAL	467 (1)	230 (1)	249 (1)
MIGRATION	62 (5)	50 (3)	49 (2)

Table 3 (continued)

	Reference	Word	Hybrid	
COMMUNITY		66 (2)	56 (4)	
INNOVATION		52 (4)	53 (3)	
ENTREPRENEURSHIP		42 (5)	42 (5)	
SOCIAL_NETWORK_ANALYSIS	217 (2)			
TRUST	92 (3)			
GENDER	78 (4)			
Reference	Citations			
GRANOVET_1973_A_1360	1502 (1)	597 (1)	625 (1)	
BURT_1992_STRUCTURAL	762 (2)	390 (2)	415 (2)	
PUTNAM_2000_BOWLING	606 (3)	241 (4)	264 (4)	
GRANOVET_1985_A_481	484 (4)	302 (3)	321 (3)	
COLEMAN_1988_A_95		222 (5)	238 (5)	
COLEMAN_1990_FDN	453 (5)			
Author	Authorships			
KILDUFF,_M	16 (4)	10 (2)	10 (2)	
HOSSAIN,_L	18 (3)	9 (4)		
FOLKE,_C		11 (1)	11 (1)	
SORENSON,_O		8 (3)	8 (3)	
DUNBAR,_RIM	19 (1)			
JACKSON,_MO	18 (2)			
BRASS,_DJ	14 (5)			
BODIN,_O		7 (5)		
ERNSTSON,_H			7 (4)	
JONES,_N			7 (5)	
(c) *Social Network Analysis*				
Publications	1187	1982	2931	2802
Year quartiles	06/09/11	07/10/11	08/10/12	07/10/12
Subject category	Classifications			
Comp. sci., artificial intelligence	131 (3)	220 (2)	417 (2)	360 (2)
Comp. sci., information systems		357 (1)	602 (1)	567 (1)
Comp. sci., theory and methods		254 (3)	457 (3)	400 (3)
Management		189 (4)	272 (5)	281 (4)
Sociology	124 (4)	169 (5)		
Comp. sci., interdisciplinary appl.			252 (4)	
Information science and library sci.				252 (5)
Zoology	99 (1)			
Anthropology	96 (2)			
Behavioral sciences	67 (5)			
Author keyword	Usages			
SOCIAL_NETWORK_ANALYSIS	194 (1)	420 (1)	1097 (1)	1105 (1)
CENTRALITY	24 (2)	36 (3)	49 (4)	53 (3)
NETWORK_ANALYSIS		53 (2)	60 (2)	68 (2)
DATA_MINE		23 (4)	53 (3)	54 (4)
CLUSTER		20 (5)	47 (5)	

Table 3 (continued)

	Reference	Word	Hybrid	
KNOWLEDGE_MANAGEMENT				37 (5)
VISUALIZATION	14 (4)			
ASSOCIATION	13 (3)			
SOCIAL_STRUCTURE	13 (5)			
Reference	Citations			
FREEMAN_1979_S_215	195 (1)	322 (3)	394 (2)	432 (2)
BORGATTI_2002_UCINET	182 (2)	307 (4)	375 (3)	405 (3)
HANNEMAN_2005_INTRO	81 (4)	128 (5)	185 (5)	195 (5)
WASSERMA_1994_SOCIAL		1817 (1)	931 (1)	1068 (1)
SCOTT_2000_SOCIAL		361 (2)	299 (4)	320 (4)
SCOTT_1991_SOCIAL	93 (3)			
FREEMAN_1977_S_35	59 (5)			
Author	Authorships			
KAZIENKO,_P		24 (1)	23 (1)	22 (1)
LEYDESDORFF,_L	13 (1)		19 (2)	20 (2)
PARK,_HW			17 (3)	17 (3)
MUSIAL,_K		19 (2)		
VALENTE,_TW		17 (3)		
BUTTS,_CT		16 (4)		
HOSSAIN,_L				16 (4)
SNIJDERS,_TAB		16 (5)		
BRANDES,_U			14 (4)	
CHEN,_HC				14 (5)
DOREIAN,_P	12 (3)			
LEE,_PC			12 (5)	
SUEUR,_C	12 (2)			
REYNOLDS,_RG	10 (4)			
VARGAS-QUESADA,_B	9 (5)			
(d) *Network Science*				
Publications	2893	3508	4066	
Year quartiles	2008/2010/2012	2007/2010/2012	2007/2010/2012	
Subject category	Classifications			
Comp. sci., information sys.	655 (2)	597 (1)	747 (2)	
Comp. sci., theory and meth.	636 (1)	520 (2)	679 (1)	
Comp. sci., artificial intel.	455 (3)	421 (3)	568 (3)	
Engineering, electrical and el.	435 (5)	345 (4)	435 (4)	
Physics, multidisc.	257 (4)		237 (5)	
Comp. sci., interdisc. appl.		265 (5)		
Author keyword	Usages			
COMPLEX_NETWORK	145 (1)	87 (1)	127 (1)	
SMALL_WORLD	50 (4)	44 (4)	49 (5)	
AGENT_BASE_MODEL		77 (2)	82 (2)	
CLUSTER	55 (5)		55 (4)	
SIMULATION		43 (5)	50 (3)	

Table 3 (continued)

	Reference	Word	Hybrid
SOCIAL_NETWORK_ANALYSIS	172 (2)		
COMMUNITY_DETECTION	75 (3)		
TRUST		58 (3)	
Reference	Citations		
WATTS_1998_N_440	777 (1)	441 (2)	641 (1)
BARABASI_1999_S_509	737 (2)	439 (1)	602 (2)
NEWMAN_2003_S_167	513 (3)	263 (4)	405 (3)
ALBERT_2002_R_47	442 (4)	278 (3)	382 (4)
WASSERMA_1994_SOCIAL		390 (5)	481 (5)
GIRVAN_2002_P_7821	325 (5)		
Author	Authorships		
NEWMAN,_MEJ	25 (2)	17 (5)	24 (2)
CHRISTAKIS,_NA	31 (1)		31 (1)
SNIJDERS,_TAB		22 (1)	24 (3)
JACKSON,_MO		19 (2)	23 (4)
WANG,_L	19 (5)		19 (5)
WU,_B	21 (4)		
FRANKS,_DW	20 (3)		
SANCHEZ,_A		17 (3)	
SANTOS,_FC		16 (4)	
(e) *Computational Social Science*			
Publications	5591	6152	5954
Year quartiles	2009/2011/2012	2009/2011/2012	2009/2011/2012
Subject category	Classifications		
Comp. sci., information sys.	1538 (1)	1709 (1)	1645 (1)
Comp. sci., theory and meth.	1261 (2)	1467 (2)	1396 (2)
Eng., electrical & electronic	914 (3)	1099 (3)	1051 (3)
Comp. sci., artificial intel.	757 (4)	793 (5)	735 (5)
Telecommunications	537 (5)	673 (4)	638 (4)
Author keyword	Usages		
FACEBOOK	304	306	322
SOCIAL_NETWORK_SITE	307	277	295
WEB_2.0	272	271	283
SOCIAL_MEDIA	232	250	255
INTERNET	244	229	239
Reference	Citations		
BOYD_2008_J_210	293 (1)	243 (1)	275 (1)
ELLISON_2007_J_1143	234 (2)	187 (2)	212 (2)
O'REILLY_2005_WHAT	158 (3)	130 (5)	144 (3)
BOYD_2007_J	135 (4)		122 (4)
ELLISON_2007_J	112 (5)		113 (5)
WASSERMA_1994_SOCIAL		295 (3)	
GRANOVET_1973_A_1360		252 (4)	

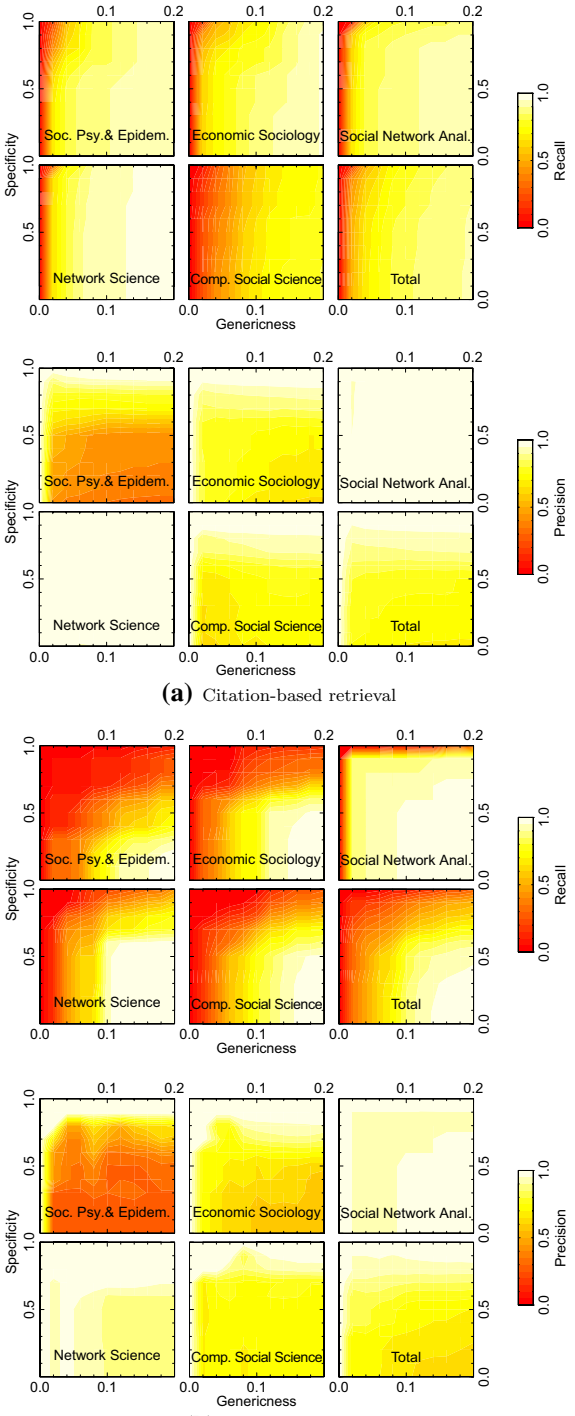
Table 3 (continued)

	Reference	Word	Hybrid
Author	Authorships		
ZHANG,_J	18 (4)	17 (2)	17 (2)
JUNG,_JJ	17 (3)	15 (3)	15 (3)
TURBAN,_E	15 (5)	14 (5)	14 (5)
ZHANG,_Y		17 (1)	17 (1)
LEE,_S		15 (4)	15 (4)
MORENO,_MA	26 (1)		
CHRISTAKIS,_DA	19 (2)		

In five subtables, subseeds are described for reference, word, and hybrid (reference/word) coupling. Each is described by its size in publications, publication year quartiles, top 5 subject categories (journal classes in the Web of Science), and top 5 facts. Values are the numbers of selections k, and ranks are given in brackets. Rankings are based on $tf * idf$ scores where $tf = k$ is the number of selections in a subseed and $idf = \log(1/K)$ is the logarithm of the inverse fraction of selecting publications in the seed. Clusters for different coupling methods are quite robust in terms of fact rankings. Interpretation and comparison of subject categories and facts allows a coherent labeling of five subseeds (shown on this and the following four pages): (a) Social Psychology and Epidemiology (SPE), (b) Economic Sociology (ES), (c) Social Network Analysis (SNA), which is split into two for reference coupling, (d) Network Science (NS), and (e) Computational Social Science (CSS). Subseeds are listed by median publication year. Author names are not yet disambiguated in this step

(relevant or not). The process can be retraced by studying the Supplementary Information (Section 2) which gives 15 examples for each class. Here, it gets clear why the boundary set should not contain publications that are largely off topic. If there were, it would be obvious if they should belong to the field or not, but such a classification is only of limited use. Since the goal was to define a science of social networks, not a sociological network science, I ruled publications inside or relevant when they are truly relational and outside or irrelevant when the NETWORK concept is used metaphorically or in non-social contexts. Publications about engineering networked social systems were ruled inside when they are not purely about issues of implementation. Given this classification, specificity $\upsilon_{j,s}$ is the fraction of publications retrieved by a fact that are relevant (as judged via the sample Ω').

Figure 4 depicts the efficiency and accuracy of the retrieval procedure and how the genericness and specificity parameters influence its recall and precision for the unpartitioned seed and broken down to the five subseeds. I had to introduce an upper limit for genericness of 0.2 to reduce manual labour in using the Web of Science online interface. The plots reveal two things. First, recall is higher for citation-based retrieval, parameter settings being equal. Reference cores are more generic or, put differently, at similar genericness, lexical retrieval is associated with a lower recall because language use is relatively imprecise. Second, idiosyncrasies of subfields point at differences of ideational closure or cultural coherence (Fuchs 2001, p. 55). Social Network Analysis has a very generic small core, few references and words suffice to retrieve a large fraction of relevant publications. Accordingly, the word SOCIAL_NETWORK_ANALYSIS is used by 1105 or 39% of the publications. 38% cite the reference WASSERMA_1994_SOCIAL (cf. Table 3c). Social Psychology and Epidemiology is at the other extreme. Lexical retrieval is either good regarding recall or precision, but not both. Only 8% of the publications use the top-ranked SOCIAL_SUPPORT and only 5% cite the top-ranked



(a) Citation-based retrieval

(b) Lexical retrieval

Fig. 4 Recall and precision of publication retrieval. Recall and precision of the subcores and the total core depend on the efficiency (genericness parameter) and accuracy (specificity parameter) of the delineation procedure. For the example of Social Psychology (SP), the subcore with a genericness of up to 10% and a specificity of at least 50% recalls 89% of the relevant publications at a precision of 43%. For lexical retrieval and the same parameters, recall is 43% at a precision of 29%. Values for the total are obtained by treating the whole field like a subfield

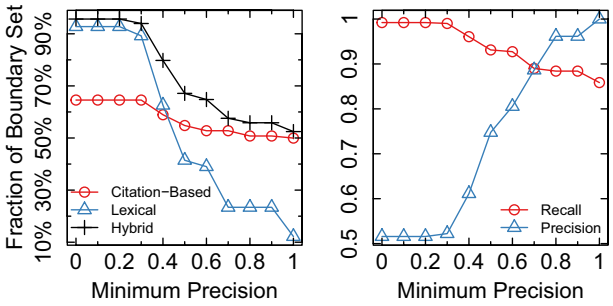


Fig. 5 Effects of boundary confidence choice. (Left) The fraction of publications retrieved from the boundary set Ω decreases with increasing minimum precision Π, but less so for citation-based retrieval. The black trend resembles the set union as the result of hybrid retrieval. (Right) Opposing trends of overall recall and precision according to hybrid retrieval of the classified sample Ω'

`BERKMAN_1979_A_186` (Table 3a). These ideosynchrasies highlight the need for a subfield-specific delineation procedure.

Assembling the field

In this third step, the second iteration E of the field is assembled from the subcores B_s. When delineating SNS using one parameter setting, the problem that a particular setting entails varying recall and precision of the subcores shows as follows. For citation-based retrieval, setting $\Upsilon = 0.5$ ($\Psi = 0.1$) results in an overall precision of 0.76 but, for Social Psychology, of 0.43. This is a direct effect of the boundary problem, and the proposed solution is to decide on a justifiable fuzziness of the boundary. In that case, the researcher must decide how the assembly parameter Π should determine a universal setting of Ψ and Υ. Demanding that Π guarantees a minimum recall for all subfields entails an increase of false positives (an increase of retrieved publications that are irrelevant) as the assembly parameter increases. As a result, the boundary will be more fuzzy. Demanding that Π guarantees a minimum precision for all subfields entails an increase of false negatives (an increase of relevant publications that are not retrieved) as the assembly parameter increases. As a result, the boundary will be less fuzzy. As I want the boundary to contain few irrelevant publications, I prioritize on accuracy and condition the genericness and specificity parameters on a minimum precision Π that should be achieved for all subfields. This parameter resembles the confidence that the boundary separates irrelevant from relevant publications. Given Π, those parameters Ψ and Υ are deduced which maximize overall recall.

Figure 5 reports effects of a given confidence in the boundary. The left plot reports the fraction of publications retrieved from the boundary set Ω. Different upper bounds are visible. For citation-based retrieval, no more than 65% of 44,308 papers can be retrieved as $\Psi \leq 0.2$ sets a technical limit. Lexical cores can retrieve 93% at low minimum precision but the fraction quickly drops when subfield boundaries are required to be less fuzzy. The advantage of hybrid retrieval, the set union from the two approaches, is that it balances the high specificity/low genericness of citation-based retrieval and the low specificity/high genericness of lexical retrieval. When the boundary from hybrid retrieval is required to be perfectly precise ($\Pi = 1$), then the field will consist of about 23,000 publications. But when 80% of the publications are allowed to be irrelevant ($\Pi = 0.2$), then the field will be about 35,000 publications strong. The right plot of Fig. 5 reports the efficiency and

Table 4 Parameters for field assembly

Retrieval	Π	Ψ	Υ	Recall	Precision
Citation-based	0.0, … , 0.3	0.20	0.0	0.92 (0.06)	0.71 (0.24)
	0.4	0.20	0.3	0.92 (0.07)	0.74 (0.21)
	0.5	0.20	0.6	0.92 (0.06)	0.82 (0.15)
	0.6,...,0.7	0.20	0.7	0.91 (0.07)	0.89 (0.09)
	0.8, … , 0.9	0.20	0.9	0.90 (0.07)	0.97 (0.02)
	1.0	0.20	1.0	0.89 (0.07)	1.00 (0.00)
Lexical	0.0, … , 0.2	0.20	0.0	0.99 (0.01)	0.68 (0.24)
	0.3	0.20	0.4	0.94 (0.09)	0.68 (0.23)
	0.4	0.20	0.6	0.86 (0.14)	0.75 (0.19)
	0.5	0.20	0.7	0.71 (0.19)	0.82 (0.14)
	0.6	0.18	0.7	0.71 (0.18)	0.83 (0.14)
	0.7, … , 0.9	0.20	0.9	0.51 (0.24)	0.98 (0.02)
	1.0	0.20	1.0	0.26 (0.12)	1.00 (0.00)

The assembly parameter Π is chosen to specify the minimum precision of all subcores. Parameters Ψ (genericness) and Υ (specificity) are deduced by maximizing overall recall. This setting corresponds to an average recall and precision (standard deviation for the five subfields in brackets). For citation-based retrieval and a mimimum precision of $\Pi = 0.8$, the best overall recall is obtained for $\Psi = 0.20$ and $\Upsilon = 0.9$. For lexical retrieval and the same assembly parameter, Ψ and Υ are the same, precision is similar but recall is lower

Table 5 Number of facts used for retrieval

	References	Citations	Words	Usages
Social Psychology and Epidem.	4685	5	42	33
Economic Sociology	3443	4	90	8
Social Network Analysis	1004	7	15	20
Network Science	1162	8	76	12
Computational Social Science	3583	4	35	34

For a genericness of $\Psi = 0.2$, the five subfields are retrieved via different numbers of facts. Ψ also translates to different absolute minima of selections

accuracy of hybrid retrieval conditional on Π. Recall is at a satisfactorily high level for the whole range of minimum precision. At this point, I set $\Pi = 0.8$ because a lower value would increase the overall fraction of false positives to over 10%. Table 4 lists the parameters and the average subfield recall and precision that can be achieved for a given minimum precision. It reveals that the parameters corresponding to $\Pi = 0.8$ are $\Psi = 0.2$ and $\Upsilon = 0.9$ and that the confidence that the subfield boundaries separate irrelevant from relevant publications is 97% on average regarding references to concept symbols and 98% regarding language use.

Table 5 states that the subfields contribute different numbers of facts to the overall retrieval core and that a genericness of 20% translates to different absolute selection thresholds. While Computational Social Science's core references are cited at least four times, an absolute threshold for Network Science would have to be eight. This demonstrates that the original method of not distinguishing subfields is only applicable to network domains

Table 6 Core description and sourcing

Mapping	SOCIAL_				
	SOCIAL_ SUPPORT	SOCIAL_ CAPITAL	NETWORK_ ANALYSIS	COMPLEX_ NETWORK	SOCIAL_ MEDIA
	SPE	ES	SNA	NS	CSS
	Core statistics				
References	2914	3338	981	1143	2768
Article fraction	73%	55%	71%	79%	46%
Chapter fraction	7%	6%	5%	3%	19%
Field closure	30%	23%	38%	29%	23%
Subfield closure	21%	15%	15%	21%	9%
	Sourcing rates				
Articles	93%	90%	91%	87%	78%
Chapters	3%	11%	14%	22%	15%
Total	68%	50%	66%	70%	38%

Clustering the field E results in five communities which are labeled by their most used word (top row). Labels map to the subseeds of Table 3. The first data row (core statistics) gives the size of each community's cited core Γ_s for a genericness of $\Psi = 0.2$. The core of the SOCIAL_SUPPORT community which maps to the seed community Social Psychology and Epidemiology (SPE) consists of 2914 references. 73% of these are articles and 7% are chapters according to the definition given in the Supplementary Information (Section 1). The rest are books. Field closure is the fraction of the cited articles that are publications in the field E. 30% of the SOCIAL_SUPPORT core references are themselves contained in SNS. Subfield closure is the fraction of the cited articles that are publications in the subfield E'_s. 21% of the core references are themselves contained in the SOCIAL_SUPPORT community. Sourcing rates tell which fraction of a community's core could successfully be found in the Web of Science database and added to the field set E. Rates are generally high for articles. Chapters include conference proceedings, some of which are included in the database

whose subfields do not have varying selection practices or differ in size. Using the deduced parameter setting, the five subcores B_s (per practice) are defined and the subfields E_s which select these subcores are retrieved. The set union of E_s is the second iteration E of SNS and consists of 24,748 publications. This is slightly larger than the size of the seed from which the delineation procedure has, until here, removed irrelevant publications and to which it has added relevant ones from the boundary set.

Extending the field

In this fourth step, the second iteration E of the field is extended by adding to it its most cited references. This step is supposed to reconstruct more complete citation paths. The procedure calls for partitioning E into subfields E'_s. Louvain community detection in the network of publications (coupled through references and words) expectedly results in five communities again (modularity $Q = 0.12$). The subfields E'_s they represent can be meaningfully mapped to subseeds A_s via the most used word (Table 6). Clustering consensus from ten runs is 99% instead of 92% for the seed (cf. Table 2). Cited cores for the field are also somewhat smaller than for the seed (cf. Table 5). Both results are evidence that the delineation procedure has resulted in more compact subfields and less fuzzy subfield boundaries.

Next, the subcores Γ_s are identified, now only using the genericness parameter $\Psi = 0.2$ (the same value as for defining the subcores B_s). Unlike the subcores B_s, the subcores Γ_s

need not be specific since scholarly fields are always open to some extent—they also cite core references outside their own boundaries. By counting the fraction of cited references in subcores Γ_s that are also contained in the field E and in the subfield E'_s, I determine the extent to which the field and its subfields are closed. Books are not covered by the Web of Science products I had accessed. Consequently, founding books like Moreno's *Who Shall Survive?* (1934) can only show up as cited references, not as citing publications. Social Network Analysis is most self-contained with respect to the whole field: 38% of its core references are themselves publications in SNS, mirroring its role as the methodological power house of the field. With respect to subfields, Social Psychology and Epidemiology (SPE) and Network Science (NS) are most closed. Computational Social Science (CSS), the youngest subfield, least cites its own publications. As expected for a subfield rooted in computer science, it cites a large fraction of conference proceedings articles (book chapters).

Table 6 further reports that only 38% of CSS's 2768 core references could be identified in the database—could be sourced—and added to the field. Its article sourcing rate is smallest, too. NS has the largest sourcing rate overall (70%) and for chapters (22%). 93% of SPE's cited articles were successfully added to the field.[5] In total, 4965 core references were added to the field. In the resulting set, the extended field Z, I removed some publications or references to prevent meaningless results, artifacts, or the failure of algorithms.[6]

The final dataset Z, the third iteration of the field, consists of 25,760 publications (journal and conference proceedings articles). Following the disambiguation of author names (Supplementary Information, Section 1.1), 45,580 author identities remain that relate to publications in 68,227 authorships. 574,036 distinct references are selected in 1,125,321 citations (180,861 to publications in SNS). Following the removal of general science language (Supplementary Information, Section 1.3), 23,026 words (occurring in title, abstract, or as author keywords) are used in 201,608 selections. These entities and relationships are displayed in Fig. 6. The dataset is publicly made available (Lietz 2019) and can be explored online (Lietz 2020).

Description of the final dataset

The earliest publication in SNS is Hanifan's "The rural school community center" from 1916 (`HANIFAN_1916_A_130`) because it is often cited as one of the first occurrences of the `SOCIAL_CAPITAL` concept. From then on, the field grows continuously with a slight tendency for superexponential growth, as can be seen in the top plot of Fig. 7. It also shows that subfields came to exist at different points in time and exhibit phases of accelerating and decelerating growth. To obtain the final subfields, the hybrid publication graph representing Z is once again clustered using Louvain community detection. Table 7 is a description of the five subfields. Labels still match those of the seed very

[5] 9142 unique cited references were chosen for extending the field. Publication identifiers for references published not earlier than 1980 were queried using the bibliometric database of the German Competence Centre for Bibliometrics (www.bibliometrie.info). Heavily cited references that could not be found as well as references published before 1980 were queried using the Web of Science online interface www.webofknowledge.com. The primary search criterion was the doi, the secondary criteria were the tagged meta data.

[6] 61 publications were removed because they did not have unique matchkeys. Five articles with an `ANONYMOUS` author were removed. Furthermore, I removed citations from a publication to a reference with the identical matchkey, references with Chinese letters, and references without a cited author, source name, or publication year. Finally, all publications published after 2012 were removed because those years were not completely covered in the database.

Fig. 6 Entity relationship diagram of the final dataset. Tables with primary keys (PK) contain entities (e.g., publications) and their attributes (e.g., the time it was published). Tables that only contain foreign keys (FK) are relational tables that can be directly used for network construction

well (cf. Table 3). The consensus of detecting these communities is 0.91, i.e., extending the field has reduced the consensus from field clustering; boundaries are fuzzier again. Modularity is low ($Q = 0.13$) because of very high density ($D = 0.57$).

Some assignments of publications to subfields are counterintuitive. For example, Heider's article on balance theory (`HEIDER_1946_J_107`) is not in Social Psychology but in Economic Sociology, together with Cartwright and Harary's graph theoretical generalization (`CARTWRIG_1956_P_277`) as well as foundational works of the Harvard school, like Granovetter's "The strength of weak ties" (`GRANOVET_1973_A_1360`) and White et al.'s article on blockmodeling (`WHITE_1976_A_730`). This makes sense because these papers belong to the sociometry tradition initiated by Moreno.

The importance of fractional selection counting in the construction of publication similarity scores is once more demonstrated by the average number of references per publication which is a characteristic score for each subfield, depicted in Fig. 7. The fact that an average paper in Economic Sociology cites almost twice as many references in 2010 than an average paper in Computational Social Science means that a citation in the latter subfield is twice as valuable. Normalized citation counts k^{N} account for such differences but are still affected by the size of the respective subfield. Publication fractions K account for size differences but not for different citation practices. Only citation fractions K^{N} are comparable across subfields (Table 7). The reference `WASSERMA_1994_SOCIAL` and word `SOCIAL_NETWORK_ANALYSIS` are about ten times more common in Social Network Analysis than the top reference `O'REILLY_2005_WHAT` in Computational Social Science or the top word `COMPLEX_NETWORK` in Network Science.

The average number of words per publication exhibits a marked jump in 1990 because that year the database producers started including abstracts and author keywords in the Web of Science database. The average number of authors per publication is constantly increasing since the 70s, marking the decade when the field started becoming a "big science" (Price 1986) where knowledge production in teams is increasingly important (Wuchty et al.

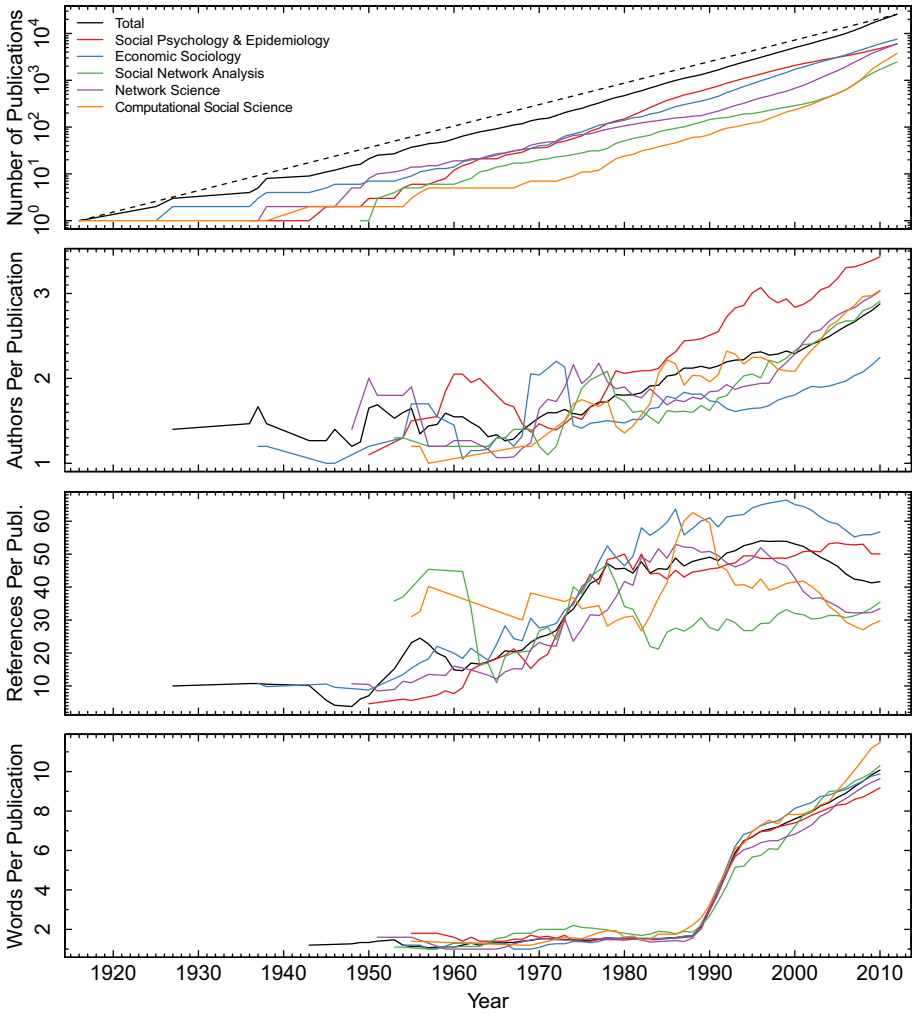


Fig. 7 Field growth and statistics. Curvature of the number of publications over time (compared to the broken line representing purely exponential growth) signals slight superexponential growth. The other curves depict the average number of facts selected in a publication per year. Words subsume words in titles and abstracts as well as author keywords

2007). There are differences, however. Economic Sociology is much less a team science than Social Psychology & Epidemiology.

Figure 8 unveils that SNS is well described by power law size distributions for authorship (Lotka's Law, Lotka 1926) and citation (Price 1976). There is also evidence for Zipf's Law (Zipf 2012 [1949]). Even though the word usage distribution is not plausibly fit by a pure power law, all subfields except Economic Sociology are plausibly fitted by Zipf exponents $\hat{\alpha}_{\text{wrd}} \approx 2$ (Lietz 2016, Table 3.8). Finally, Fig. 1 displays the cores of the field's

Table 7 Subfields in Social Network Science

(a) *Social Psychology and Epidemiology*					
Publications	5945				
Year quartiles	1997/2005/2010				
Subject category	Classifications				
Public, env. and occup. health	1049				
Sociology	757				
Psychology, multidisc.	497				
Psychology, social	497				
Psychology, developmental	478				
Word	k	k^N	K	K^N	l
SOCIAL_SUPPORT	1397	291.55	23.50%	4.90%	37
FRIEND	978	126.89	16.45%	2.13%	34
COMMUNITY	636	77.66	10.70%	1.31%	29
FRIENDSHIP	358	48.91	6.02%	0.82%	32
SOCIAL_SUPPORT_NETWORK	277	51.28	4.66%	0.86%	31
SOCIAL_RELATIONSHIP	306	43.16	5.15%	0.73%	28
LONGITUDINAL_STUDY	240	31.53	4.04%	0.53%	26
EMOTIONAL_SUPPORT	215	26.53	3.62%	0.45%	23
LONELINESS	205	36.15	3.45%	0.61%	32
SOCIAL_CAPITAL	341	38.44	5.74%	0.65%	18
Reference	k	k^N	K	K^N	l
BERKMAN_1979_A_186	384	9.25	6.46%	0.16%	33
COHEN_1985_P_310	361	7.46	6.07%	0.13%	27
HOUSE_1988_S_540	289	6.57	4.86%	0.11%	24
COBB_1976_P_300	272	6.34	4.58%	0.11%	34
RADLOFF_1977_A_385	261	6.10	4.39%	0.10%	31
FISCHER_1982_DWELL	291	7.26	4.89%	0.12%	31
GRANOVET_1973_A_1360	472	10.03	7.94%	0.17%	36
ROOK_1984_J_1097	183	3.74	3.08%	0.06%	29
PUTNAM_2000_BOWLING	266	5.67	4.47%	0.10%	12
HOUSE_1981_WORK	183	4.17	3.08%	0.07%	31
Author	k	k^N	K	K^N	l
LATKIN,_CARL	68	17.78	1.14%	0.30%	17
BERKMAN,_LISA	42	13.95	0.71%	0.23%	24
KAWACHI,_ICHIRO	31	7.71	0.52%	0.13%	14
LITWIN,_HOWARD	28	22.00	0.47%	0.37%	15
ANTONUCCI,_TONI	24	8.33	0.40%	0.14%	17
DUNBAR,_ROBIN_I_M	22	11.08	0.37%	0.19%	10
VALENTE,_THOMAS	22	7.13	0.37%	0.12%	11
CELENTANO,_DAVID	20	2.96	0.34%	0.05%	16
COHEN,_SHELDON	19	6.99	0.32%	0.12%	17
KRAUSE,_NEAL	19	12.23	0.32%	0.21%	15
(b) *Economic Sociology*					
Publications	7554				
Year quartiles	2001/2007/2010				

Table 7 (continued)

Subject category	Classifications				
Management	1783				
Business	1293				
Sociology	1357				
Geography	665				
Economics	628				
Word	k	k^N	K	K^N	l
SOCIAL_CAPITAL	1210	147.52	16.02%	1.95%	22
ORGANIZATIONAL	614	74.90	8.13%	0.99%	29
INNOVATION	582	68.71	7.70%	0.91%	30
COMMUNITY	896	94.89	11.86%	1.26%	23
OPPORTUNITY	555	58.94	7.35%	0.78%	24
TRUST	538	60.01	7.12%	0.79%	24
AGENCY	324	39.07	4.29%	0.52%	23
GOVERNANCE	291	31.21	3.85%	0.41%	22
ENTREPRENEUR	264	26.75	3.49%	0.35%	21
INFORMAL	284	29.83	3.76%	0.39%	28
Reference	k	k^N	K	K^N	l
GRANOVET_1973_A_1360	1143	23.01	15.13%	0.30%	38
GRANOVET_1985_A_481	841	16.26	11.13%	0.22%	27
BURT_1992_STRUCTURAL	884	17.45	11.70%	0.23%	21
COLEMAN_1988_A_10.1086/228943	799	15.71	10.58%	0.21%	22
COLEMAN_1990_FDN	586	10.88	7.76%	0.14%	18
PUTNAM_1993_MAKING	505	10.27	6.69%	0.14%	18
PUTNAM_2000_BOWLING	521	11.43	6.90%	0.15%	13
UZZI_1997_A_35	374	6.65	4.95%	0.09%	16
NAHAPIET_1998_A_242	371	7.56	4.91%	0.10%	14
PORTES_1998_A_1	370	6.83	4.90%	0.09%	14
Author	k	k^N	K	K^N	l
FOLKE,_CARL	15	4.71	0.20%	0.06%	8
SORENSON,_OLAV	13	7.33	0.17%	0.10%	7
GULATI,_RANJAY	13	7.83	0.17%	0.10%	8
CROSS,_ROB	13	5.12	0.17%	0.07%	5
EISENHARDT,_KATHLEEN	13	8.00	0.17%	0.11%	11
YEUNG,_H_W_C	12	9.90	0.16%	0.13%	8
STUART,_TOBY	12	6.33	0.16%	0.08%	10
BURT,_RONALD_S	13	10.42	0.17%	0.14%	11
MCEVILY,_BILL	11	4.33	0.15%	0.06%	8
PORTES,_A	11	6.50	0.15%	0.09%	8
(c) *Social Network Analysis*					
Publications	2459				
Year quartiles	2006/2009/2011				
Subject category	Classifications				
Computer science, information systems	500				
Computer science, theory and methods	350				

Table 7 (continued)

Computer science, artificial intelligence	305				
Information science and library science	279				
Computer science, interdisc. applications	241				
Word	k	k^N	K	K^N	l
SOCIAL_NETWORK_ANALYSIS	1846	271.81	75.07%	11.05%	34
NETWORK_ANALYSIS	248	32.97	10.09%	1.34%	26
SOCIAL_NETWORK_ANALYSIS_SNA	190	19.89	7.73%	0.81%	12
CENTRALITY	205	24.90	8.34%	1.01%	23
COMMUNITY	304	31.40	12.36%	1.28%	18
COLLABORATION	179	16.94	7.28%	0.69%	14
SOCIAL_STRUCTURE	159	32.74	6.47%	1.33%	38
NETWORK_STRUCTURE	136	11.82	5.53%	0.48%	16
WEB	124	11.31	5.04%	0.46%	11
DATA_MINE	83	8.96	3.38%	0.36%	11
Reference	k	k^N	K	K^N	l
WASSERMA_1994_SOCIAL	1034	51.37	42.05%	2.09%	19
FREEMAN_1979_S_215	407	14.37	16.55%	0.58%	28
BORGATTI_2002_UCINET	362	12.25	14.72%	0.50%	10
SCOTT_2000_SOCIAL	286	10.81	11.63%	0.44%	11
HANNEMAN_2005_INTRO	156	5.19	6.34%	0.21%	8
SCOTT_1991_SOCIAL	154	9.55	6.26%	0.39%	18
BURT_1992_STRUCTURAL	207	5.10	8.42%	0.21%	18
GRANOVET_1973_A_1360	250	7.01	10.17%	0.29%	30
FREEMAN_1977_S_35	112	3.72	4.55%	0.15%	19
DE_2005_EXPLORATORY	88	4.14	3.58%	0.17%	7
Author	k	k^N	K	K^N	l
LEYDESDORFF,_LOET	22	14.17	0.89%	0.58%	11
KAZIENKO,_PRZEMYSLAW	20	7.50	0.81%	0.31%	6
GLOOR,_PETER	16	5.76	0.65%	0.23%	7
BRANDES,_ULRIK	14	5.61	0.57%	0.23%	10
PARK,_HAN_WOO	14	6.25	0.57%	0.25%	6
DOREIAN,_PATRICK	13	7.83	0.53%	0.32%	11
CARLEY,_KATHLEEN	13	5.39	0.53%	0.22%	9
HOSSAIN,_LIAQUAT	12	5.00	0.49%	0.20%	5
CHEN,_CHAOMEI	9	3.46	0.37%	0.14%	6
SHNEIDERMAN,_BEN	9	3.40	0.37%	0.14%	6
(d) *Network Science*					
Publications	6031				
Year quartiles	2005/2009/2011				
Subject category	Classifications				
Computer science, theory and methods	996				
Computer science, information systems	1073				
Computer science, artificial intelligence	881				
Engineering, electrical and electronic	799				
Physics, multidisc.	531				

Table 7 (continued)

Word	k	k^N	K	K^N	l
COMPLEX_NETWORK	594	67.25	9.85%	1.12%	15
COMMUNITY	655	72.23	10.86%	1.20%	19
NETWORK_STRUCTURE	367	43.22	6.09%	0.72%	25
CONNECTIVITY	315	35.81	5.22%	0.59%	20
SMALL_WORLD	279	33.79	4.63%	0.56%	18
AVERAGE	332	39.50	5.50%	0.65%	20
SCALE_FREE_NETWORK	256	29.76	4.24%	0.49%	13
EVOLVE	319	33.97	5.29%	0.56%	17
COMMUNITY_STRUCTURE	258	26.59	4.28%	0.44%	14
COOPERATION	302	31.12	5.01%	0.52%	20
Reference	k	k^N	K	K^N	l
WATTS_1998_N_440	1101	46.12	18.26%	0.76%	14
BARABASI_1999_S_509	1018	40.41	16.88%	0.67%	13
ALBERT_2002_R_47	708	27.94	11.74%	0.46%	11
NEWMAN_2003_S_167	675	26.54	11.19%	0.44%	10
WASSERMA_1994_SOCIAL	623	20.08	10.33%	0.33%	18
GIRVAN_2002_P_7821	355	13.49	5.89%	0.22%	10
PASTOR-S_2001_P_3200	275	9.32	4.56%	0.15%	12
AMARAL_2000_P_11149	273	8.74	4.53%	0.14%	13
WATTS_1999_SMALL	290	12.03	4.81%	0.20%	14
STROGATZ_2001_N_268	255	8.65	4.23%	0.14%	12
Author	k	k^N	K	K^N	l
NEWMAN,_M_E_J	43	26.73	0.71%	0.44%	12
BARABASI,_ALBERT	41	14.14	0.68%	0.23%	12
VESPIGNANI,_ALESSANDRO	26	8.79	0.43%	0.15%	12
LATORA,_VITO	25	6.90	0.41%	0.11%	11
NOWAK,_MARTIN	23	7.82	0.38%	0.13%	12
PACHECO,_JORGE	22	7.32	0.36%	0.12%	8
EGUILUZ,_VICTOR	22	7.10	0.36%	0.12%	10
SNIJDERS,_TOM_A_B	23	11.48	0.38%	0.19%	15
KLEINBERG,_JON	22	8.90	0.36%	0.15%	11
MORENO,_YAMIR	22	5.72	0.36%	0.09%	9

(e) *Computational Social Science*

Publications	3771				
Year quartiles	2008/2010/2011				
Subject category	Classifications				
Computer science, information systems	1315				
Computer science, theory and methods	982				
Engineering, electrical and electronic	733				
Computer science, artificial intelligence	621				
Computer science, software engineering	460				
Word	k	k^N	K	K^N	l
USER	1989	187.69	52.74%	4.98%	23
INTERNET	687	63.25	18.22%	1.68%	18

Table 7 (continued)

FACEBOOK	560	51.14	14.85%	1.36%	7
SOCIAL_NETWORK_SITE	462	44.44	12.25%	1.18%	8
WEB	516	48.70	13.68%	1.29%	16
TWITTER	395	38.20	10.47%	1.01%	5
WEB_2.0	371	32.84	9.84%	0.87%	7
ONLINE_SOCIAL_NETWORK	349	35.78	9.25%	0.95%	9
BLOG	330	27.52	8.75%	0.73%	9
SOCIAL_MEDIA	317	27.26	8.41%	0.72%	8
Reference	k	k^{N}	K	K^{N}	l
O'REILLY_2005_WHAT	143	6.82	3.79%	0.18%	7
STEINFIE_2007_J_1143	104	3.30	2.76%	0.09%	5
BOYD_2007_J_210	88	3.34	2.33%	0.09%	4
ELLISON_2007_J	83	2.67	2.20%	0.07%	5
BOYD_2007_J	82	3.05	2.17%	0.08%	5
GRANOVET_1973_A_1360	171	5.64	4.53%	0.15%	21
ADOMAVIC_2005_I_734	68	3.66	1.80%	0.10%	7
DONATH_2004_B_71	67	1.81	1.78%	0.05%	7
HERLOCKE_2004_A_5	63	2.95	1.67%	0.08%	8
GOLDER_2006_J_198	57	2.77	1.51%	0.07%	7
Author	k	k^{N}	K	K^{N}	l
JUNG,_JASON_J	19	15.58	0.50%	0.41%	7
THELWALL,_MIKE	12	7.33	0.32%	0.19%	6
CARMINATI,_BARBARA	10	3.23	0.27%	0.09%	6
FERRARI,_ELENA	10	3.23	0.27%	0.09%	6
SUNDARAM,_HARI	10	3.20	0.27%	0.08%	4
PASSARELLA,_ANDREA	10	2.85	0.27%	0.08%	7
DECKER,_STEFAN	9	2.78	0.24%	0.07%	5
GOLBECK,_JENNIFER	9	6.17	0.24%	0.16%	6
LIN,_YU_RU	9	2.20	0.24%	0.06%	3
ALMEIDA,_VIRGILIO	8	1.66	0.21%	0.04%	3

In five subtables, subfields are described by their size in publications, publication year quartiles, top 5 Web of Science subject categories, and top 10 facts (ranked as described in the "Technical Appendix"). k is the number of selections, k^{N} the normalized number of selections, K the fraction of all publications, and K^{N} the fraction of all selections (all normalizations made on the subfield level). The lifetime l is the number of years in which a fact is selected at least once

three practices, all created from the same genericness threshold.[7] These graphs are filtered counterparts of the normalized fact co-selection matrices $I^{\mathrm{N}}_{\mathrm{aut}}$, $I^{\mathrm{N}}_{\mathrm{ref}}$, and $I^{\mathrm{N}}_{\mathrm{wrd}}$. As shown in Fig. 2, communities of vertices in fact-coupled transaction matrices translate to communities of edges (Ahn et al. 2010) in fact co-selection matrices. Hence, edge colors indicate

[7] Using the final dataset and for all three practices, facts with a genericness $\psi_{j,s} > 0.2$ are removed. Further filters are applied to make the plots more readable: from left to right, the top 100%, top 5%, and top 1%, respectively, of the strongest ties are kept. Graphs are then reduced to the largest bicomponent, resulting in 243, 5624, and 16 vertices, respectively. Vertex size depicts "the total contribution of [a fact] to [the publications that select it]." (Batagelj and Cerinšek 2013, p. 854). Elsewhere (Lietz 2016), I have interpreted this as the extent to which a fact catalyzes itself.

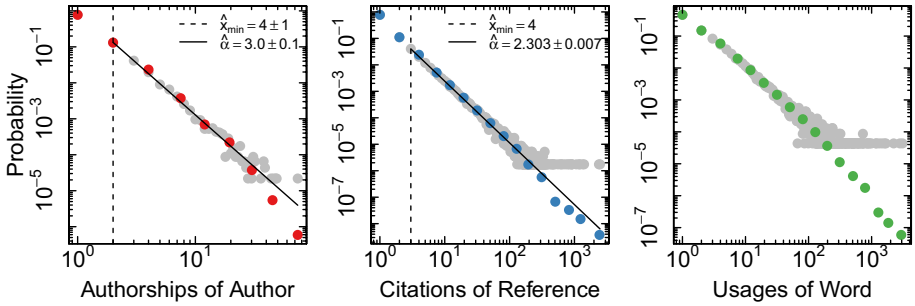


Fig. 8 Size distributions of Social Network Science. Probability density functions for authorship and citation are plausibly fit by a pure power law $p(k) \sim k^{-\alpha}$ using the maximum likelihood method (Clauset et al. 2009; Gillespie 2015). Fits are plausible if $p > 0.1$, and the scores are $p_{\text{aut}} = 0.30$ and $p_{\text{ref}} = 0.39$, respectively. Colored points result from logarithmic binning and show that extreme values also fall on the straight lines. The best power-law fit to the word usage distribution ($\hat{x}_{\text{min}} = 10 \pm 14$, $\hat{\alpha} = 2.0 \pm 0.1$) is not plausible ($p_{\text{wrd}} = 0.00$). The plots show the actual fit parameters but legends give 95% confidence intervals from bootstrapping

how facts are overlappingly co-selected in different subfields. The differences in network size mostly result from differences in concentration indicated by the decrease of exponents from $\hat{\alpha}_{\text{aut}} \approx 3$ to $\hat{\alpha}_{\text{wrd}} \approx 2$. This is mirrored in the observation that the weighted fraction K^{N} of similarly ranked facts is always (much) larger for words than for references (cf. Table 7).

Discussion and conclusion

All data is produced for a purpose, and I have presented a procedure to retrieve a research dataset representing a socio-cultural field from a corpus not necessarily created for a research purpose. The method is a development of the field delineation procedure of Zitt and Bassecoulard (2006) which departs from expert knowledge but minimizes the associated risk of expert bias by bibliometrically enhancing retrieval via a citing/cited/citing logic. By (a) mapping this logic to the mechanism of how fields reproduce themselves through positive feedback, (b) modifying it to be able to account for field heterogeneity, and (c) generalizing the routine to be able to delineate any type of field, I have proposed a sociologically enhanced information retrieval method. The reliance on a reproductive *mechanism* (Padgett and Powell 2012) effectively mitigates the risk associated with hidden assumptions because "the field writes the query." The risk is further reduced by modifying the way expert knowledge is used. Whereas, in the original method, expert knowledge is used to define a precise seed set of transactions, in the method proposed here, it is used to decide if transactions in a candidate set should be inside or outside the field boundary. As my method requires inductive, not deductive, reasoning, experts are enabled to learn and identify, and transcend, their priors (Arthur 1994).

One may ask if this gain is worth a complicated procedure with three parameters, a non-deterministic sub-procedure (clustering), and some manual classification. In the case presented here, why not simply use the SN17 dataset (Maltseva and Batagelj 2019) described in the introduction? The answer lies in the sociological boundary problem that any delineation is a construction. The SN17 dataset and the one discussed here serve different research purposes. If one is fine with having many publications in the corpus that use "social

networks" metaphorically, then SN17 is fine. Parameter-free methods as applied to delineate SN17 tend to be black boxes. But all delineations being constructions means that the steps made in field delineation are already the first steps of field analysis. This should be visible in this paper. Therefore, if one wants to have control over the boundary, my method may be an option. Then, the third parameter Π—here, chosen to be the minimum precision—serves as a goodness-of-boundary measure. That said, the SN17 dataset can also be retrieved with the method described here. First, the seed and boundary sets are created as the same set using the same `SOCIAL NETWORK*` search term and the boundary sample is coded as described. Second, the genericness parameter Ψ is set to 1 to retrieve publications via all facts, and the specificity parameter Υ is set to 0 to also use all facts for retrieval. Π is then not a retrieval parameter anymore, but a characterization of boundary fuzziness.

Still, evaluations are necessary. Since the mechanistic approach delineates fields in an organic way, certain statistical properties of dynamic systems are expected and can be used for a soft kind of evaluation, namely exponential growth and power law size distributions (Price 1986). While exponential growth is the hallmark of complex adaptive innovation systems, power laws are expected signatures because, as a functional pattern, they point at an optimization process that results in fractal structures (West 2017). Both signatures are found. The field grows slightly superexponentially, indicating that it is innovating successfully. There are also no gaps which could indicate that publications of a particular period have been missed. The field obeys Lotka's Law and exhibits a power law distribution for the citation practice. Language use statistics deserve a closer look. For the whole field, the size distribution for word usage is not plausibly fit by a power law, but for four of the five subfields, Zipf's Law holds. This leads to three conjectures. First, the field has not yet self-organized to a scale-free pattern. Second, the way natural language was processed introduced a bias. Third, delineation on the subfield level does not necessarily create a coherent whole. While the first conjecture resembles a finding, the last ones may be limitations of the method and deserve future attention.

Clear limitations exist. First, subcores were defined disregarding time, i.e., they are most effected by recent years with many publications. This recency effect can be avoided by using dynamic community detection. Not identifying subcores over time was the price of using the Web of Science and retrieving data via the online interface. To ease research, database producers could consider calls for data access where users are granted improved data access under defined terms of use. Second, abstracts and author keywords are not available for years before 1990. Since I used all author keywords as the vocabulary which is then extracted from titles and abstracts, I rely on the situation that all relevant keywords are at least used once in, or after, 1990. I think this is a fair assumption. Third, I also provided the expert knowledge when ruling candidate publications inside or outside SNS, i.e., there is no reliability check. For the reader to retrace my decisions, I present selected cases in the Supplementary Information (Section 2).

The dataset has high face validity. Subfield descriptions are robust throughout the delineation procedure (Tables 3, 7). The analysis of the dataset—not described here but in my dissertation (Lietz 2016)—reproduces results known from the previous literature, namely, roots in social psychology and graph theory, a structuralist narrative starting in the 70s, and a turn at the end of the century driven by physics (Freeman 2004; Scott 2012; Maltseva and Batagelj 2019). But the study also uncovers new insights into field dynamics, particularly regarding the paradigm shifting effects that arise when an incommensurable research style forcefully and massively enters a field. In the case of SNS, the mainstream was lastingly altered and old knowledge got more or less lost (Lietz 2016).

I conclude that the boundary constructed for SNS is a fair delineation of the field for the purpose of studying its historical evolution. The main contribution of this paper is a

sociologically enhanced information retrieval method that integrates a field model, a retrieval model, and a data model. There is indeed a benefit in importing more social science into information science (Leydesdorff and Van Den Besselaar 1997; Cronin 2008). The fact that a reproductive mechanism is at the heart of the procedure makes it principally applicable to other settings. For example, in a social media monitoring context it is typically difficult to foresee which semantic selectors (e.g., hashtags) will be used in a monitoring phase. It is much easier to define which users (e.g., politicians) are relevant (Stier et al. 2018). Future delineations of a social media monitoring corpus may be improved by starting with a user-based seed set, monitoring the emergent pattern of potential selectors, and adding/removing selectors if necessary.

Acknowledgements Open Access funding provided by Projekt DEAL. My thanks go out to Marcos Oliveira, Olga Zagovora, Indira Sen, and the three reviewers for stimulating discussions and making this manuscript more readable, to Lothar Krempel for encouraging me to make the SNS dataset publically available, and to Clarivate Analytics, the producer of the Web of Science database, for allowing me to do that. Use of the bibliometric database of the German Competence Centre for Bibliometrics, funded by the Federal Ministry of Education and Research (01PQ13001), is acknowledged.

Technical Appendix: formalism and index construction

Matrix notation Let $G = [w_{ij}]$ be a bipartite $m \times n$ *selection matrix* where w_{ij} gives the number of times that a transaction a_i, $i = 1, 2, \ldots, m$, is selecting a fact f_j, $j = 1, 2, \ldots, n$. $G^{\mathrm{T}} = [w_{ij}]^{\mathrm{T}} = [w_{ji}]$ is the *transposed selection matrix* with dimensionality $n \times m$. To control for the possibility that the average number of facts per transaction is heterogenously distributed over subfields, matrix normalization is necessary. For a selection matrix G with row indices i and column indices j, the degree $k_i = \sum_j w_{ij}$ is the number of selections made in transaction i. Batagelj and Cerinšek (2013) have generalized normalization using fractional selection counting in a matrix framework. $G^{\mathrm{N}} = \mathrm{diag}(1/\max(1, k_i))G$ is the *normalized selection matrix*. In this matrix, the weighted degree $k_i^{\mathrm{N}} = \sum_j w_{ij}^{\mathrm{N}} = 1$, i.e., the weight of a selection is inversely proportional to the number of selections per transaction.

Fact statistics With this notation in place, statistics based on degrees of facts j in selection matrices can be made (applications to bibliographic data are given in brackets):

the degree $k_j = \sum_i w_{ij}$ in the selection matrix G is the number of selections (authorships/citations/usages) of fact (author/reference/word) j;

the weighted degree $k_j^{\mathrm{N}} = \sum_i w_{ij}^{\mathrm{N}}$ in the normalized selection matrix G^{N} is the normalized number of selections (authorships/citations/usages) of fact (author/reference/word) j;

the fraction $K_j = k_j/m$ is the percentage of all transactions (publications) that select (are authored by/cite/use) fact (author/reference/word) j;

the weighted fraction $K_j^{\mathrm{N}} = k_j^{\mathrm{N}}/m$ is the percentage of all selections (authorships/citations/usages) of fact (author/reference/word) j.

Matrix projection The delineation procedure requires fields represented by sets of transactions to be partitioned into subfields. The graph to be clustered is a fact-coupled *transaction graph*, and its matrix is obtained via matrix multiplication:

$H = (GG^{\mathrm{T}}) = [x_{ik}]$ is an undirected *transaction matrix* where weights $x_{ik} \in \mathbb{N}$ are the number of facts (authors/references/words) j co-selected by (co-authoring/co-cited by/co-used by) transactions (publications) i and k;

$H^{\mathrm{N}} = (G^{\mathrm{N}}(G^{\mathrm{N}})^{\mathrm{T}}) = [x_{ik}^{\mathrm{N}}]$ is an undirected *normalized transaction matrix* where weights $x_{ik}^{\mathrm{N}} \in \mathbb{R}_{[0,1]}$ are the products of the normalized selections (authorships/citations/usages) made in transactions (publications) i and k, summed over all facts (authors/references/words) j.

H^{N} is the complementary transformation of the one described by Batagelj and Cerinšek (2013, sec. 3.4). Weights x_{ik}^{N} can be interpreted as publication similarities. The transaction matrix resembles the projection of the bipartite selection matrix to the transaction mode. The projection to the fact mode creates the matrix of a transaction-coupled *fact co-selection graph*:

$I = G^{\mathrm{T}}G = [y_{jl}]$ is a symmetric directed *fact co-selection matrix* where weights $y_{jl} \in \mathbb{N}$ are the number of transactions (publications) that co-select (are co-authored by/co-cite/co-use) facts (authors/references/words) j and l (Batagelj and Cerinšek 2013, sec. 3.2);

$I^{\mathrm{N}} = G^{\mathrm{T}}G^{\mathrm{N}} = [y_{jl}^{\mathrm{N}}]$ is a symmetric directed *normalized fact co-selection matrix* where weights $y_{jl}^{\mathrm{N}} \in \mathbb{R}_{\geq 0}$ are the normalized number of transactions (publications) that co-select (are co-authored by/co-cite/co-use) facts (authors/references/words) j and l (Batagelj and Cerinšek 2013, sec. 3.3).

I^{N} has two handy properties. First, the total weight $y_{\Sigma}^{\mathrm{N}} = \sum_j \sum_l y_{jl}^{\mathrm{N}}$ including self-loops $j = l$ equals the number of selections in the underlying selection matrix G because weights are additive. Second, the weighted degree of fact j equals the number of transactions (publications) that have selected (been authored by/cited/used) it. The reason to keep symmetric directed edges is that they can be interpreted as catalytic relations in the sense that co-selection also means co-constitution (Padgett and Powell 2012, chapter 4).

A toy selection matrix G^{N}, its mapping to the field model, and its projections to the transaction matrix H^{N} and fact matrix I^{N} is depicted in Fig. 2.

Fact genericness The delineation procedure also requires facts to be indexed by genericness and specificity scores. This indexation is made for distinct subseeds A_s represented by selection matrices G_s. To obtain genericness scores for facts in subseed A_s, facts are $tf * idf$-ranked descendingly (fact with largest score has first rank). Here, $tf = k_{j,s}$ is the degree of fact j in subseed s, and $idf = \log(1/K_j)$ where K_j is the transaction fraction of fact j in the whole seed A. Given fact ranks $r = 1, \ldots, m$, the genericness of fact j in subseed s is $\psi_{j,s} = \sum_{q=1}^{r} K_{q,s}^{\mathrm{N}}$, the cumulative sum of selection fractions. Fact rankings for subfields are created similarly.

References

Abbott, A. (2001). *Chaos of disciplines*. Chicago, IL: University of Chicago Press.

Ahn, Y. Y., Bagrow, J. P., & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, *466*(7307), 761–764. https://doi.org/10.1038/nature09182.

Arthur, W. B. (1994). Inductive reasoning and bounded rationality. *The American Economic Review*, *84*(2), 406–411.

Barabási, A. L. (2016). *Network science*. Cambridge: Cambridge University Press.

Batagelj, V., & Cerinšek, M. (2013). On bibliographic networks. *Scientometrics*, *96*(3), 845–864. https://doi.org/10.1007/s11192-012-0940-1.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *10*, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Bourdieu, P., & Wacquant, L. (1992). *An invitation to reflexive sociology*. Chicago, IL: University of Chicago Press.

Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science*, *42*(4), 233–251. https://doi.org/10.1002/(SICI)1097-4571(199105)42:4<233::AID-ASI1>3.0.CO;2-I.

Bradford, S. C. (1985 [1934]). Sources of information on specific subjects. *Journal of Information Science* *10*(4), 176–180. https://doi.org/10.1177/016555158501000407.

Brandes, U., & Pich, C. (2011). Explorative visualization of citation patterns in social network research. *Journal of Social Structure*, *12*(8), 1–19.

Breiger, R. L. (1974). The duality of persons and groups. *Social Forces*, *53*(2), 181–190. https://doi.org/10.1093/sf/53.2.181.

Callon, M., Law, J., & Rip, A. (1986). *Mapping the dynamics of science and technology: Sociology of science in the real world*. London: Macmillan.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111.

Cronin, B. (2008). The sociological turn in information science. *Journal of Information Science*, *34*(4), 465–475. https://doi.org/10.1177/0165551508088944.

Doreian, P., Batagelj, V., & Ferligoj, A. (2004). Generalized blockmodeling of two-mode network data. *Social Networks*, *26*(1), 29–53. https://doi.org/10.1016/j.socnet.2004.01.002.

Durkheim, E. (1982 [1895]). The rules of sociological method. New York, NY: Free Press.

Eck, N Jv, & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology*, *60*(8), 1635–1651. https://doi.org/10.1002/asi.21075.

Emirbayer, M. (1997). Manifesto for a relational sociology. *American Journal of Sociology*, *103*(2), 281–317.

Emirbayer, M., & Mische, A. (1998). What is agency? *American Journal of Sociology*, *103*(4), 962–1023. https://doi.org/10.1086/231294.

Flack, J. C. (2017). Coarse-graining as a downward causation mechanism. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *375*(2109), 20160338. https://doi.org/10.1098/rsta.2016.0338.

Fleck, L. (1979 [1935]). Genesis and development of a scientific fact. Chicago, IL: The University of Chicago Press.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3–5), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002.

Freeman, L. C. (2004). *The development of social network analysis: A study in the sociology of science*. Vancouver, BC: Empirical Press.

Freeman, L. C. (2011). The development of social network analysis—With an emphasis on recent events. In J. Scott & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis, chap 3* (pp. 26–39). London: SAGE.

Fuchs, S. (2001). *Against essentialism: A theory of culture and society*. Cambridge: Harvard University Press.

Fuhse, J. A. (2009). The meaning structure of social networks. *Sociological Theory*, *27*(1), 51–73. https://doi.org/10.1111/j.1467-9558.2009.00338.x.

Garfield, E. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*. New York, NY: Wiley.

Garfield, E. (2004). Historiographic mapping of knowledge domains literature. *Journal of Information Science*, *30*(2), 119–145. https://doi.org/10.1177/0165551504042802.

Garfield, E., & Sher, I. H. (1993). KeyWords Plus™—Algorithmic derivative indexing. *Journal of the American Society for Information Science*, *44*(5), 298–299. https://doi.org/10.1002/(SICI)1097-4571(199306)44:5%3C298::AID-ASI5%3E3.0.CO;2-A.

Gillespie, C. S. (2015). Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v064.i02.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367. https://doi.org/10.1023/A:1022378804087.

Glänzel, W., & Thijs, B. (2011). Using "core documents" for the representation of clusters and topics. *Scientometrics*, *88*(1), 297–309. https://doi.org/10.1007/s11192-011-0347-4.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380. https://doi.org/10.1086/225469.

Hidalgo, C. A. (2016). Disconnected, fragmented, or united? A trans-disciplinary review of network science. *Applied Network Science*, *1*(1), 6. https://doi.org/10.1007/s41109-016-0010-3.

Hummon, N. P., & Carley, K. M. (1993). Social networks as normal science. *Social Networks*, *15*(1), 71–106. https://doi.org/10.1016/0378-8733(93)90022-D.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25. https://doi.org/10.1002/asi.5090140103.

Lancichinetti, A., & Fortunato, S. (2012). Consensus clustering in complex networks. *Scientific Reports*, *2*, 336. https://doi.org/10.1038/srep00336.

Lazer, D., & Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology*, *43*(1), 19–39. https://doi.org/10.1146/annurev-soc-060116-053457.

Lazer, D., Mergel, I., & Friedman, A. (2009). Co-citation of prominent social network articles in sociology journals: The evolving canon. *Connections*, *29*(1), 43–64.

Leydesdorff, L., & Opthof, T. (2010). Normalization at the field level: Fractional counting of citations. *Journal of Informetrics*, *4*(4), 644–646. https://doi.org/10.1016/j.joi.2010.05.003.

Leydesdorff, L., & Van Den Besselaar, P. (1997). Scientometrics and communication theory: Towards theoretically informed indicators. *Scientometrics*, *38*(1), 155–174. https://doi.org/10.1007/BF02461129.

Leydesdorff, L., Schank, T., Scharnhorst, A., & Nooy, Wd. (2008). Animating the development of social networks over time using a dynamic extension of multidimensional scaling. *El Profesional de la Información*, *17*(6), 611–626.

Lietz, H. (2016). Scale-free identity: The emergence of Social Network Science. Dissertation, University of Duisburg-Essen, Faculty of Social Sciences.

Lietz, H. (2019). Social network science (1916–2012). SowiDataNet|datorium. https://doi.org/10.7802/1.1954.

Lietz, H. (2020). compsoc—Notebooks for computational sociology. Retrieved June 7, 2020 from https://github.com/gesiscss/compsoc.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences*, *16*, 317–323.

Maltseva, D., & Batagelj, V. (2019). Social network analysis as a field of invasions: Bibliographic approach to study SNA development. *Scientometrics*, *121*(2), 1085–1128. https://doi.org/10.1007/s11192-019-03193-x.

McLean, P. D. (2017). *Culture in networks*. Cambridge: Polity.

Milanez, D. H., Noyons, E., & de Faria, L. I. L. (2016). A delineating procedure to retrieve relevant publication data in research areas: The case of nanocellulose. *Scientometrics*, *107*(2), 627–643. https://doi.org/10.1007/s11192-016-1922-5.

Mogoutov, A., & Kahane, B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. *Research Policy*, *36*(6), 893–903. https://doi.org/10.1016/j.respol.2007.02.005.

Neuhaus, C., & Daniel, H. D. (2009). A new reference standard for citation analysis in chemistry and related fields based on the sections of chemical abstracts. *Scientometrics*, *78*(2), 219–229. https://doi.org/10.1007/s11192-007-2007-2.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103.

Padgett, J. F., & Powell, W. W. (2012). *The emergence of organizations and markets*. Princeton, NJ: Princeton University Press.

Page, S. E. (2015). What sociologists should know about complexity. *Annual Review of Sociology*, *41*(1), 21–41. https://doi.org/10.1146/annurev-soc-073014-112230.

Palla, G., Barabási, A. L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, *446*(7136), 664–667. https://doi.org/10.1038/nature05670.

Price, D Jd S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. https://doi.org/10.1002/asi.4630270505.

Price, D Jd S. (1986). *Little science, big science... and beyond*. New York, NY: Columbia University Press.

Schmitt, M. (2019). Felder und Netzwerkdomänen in der Wissenschaft: Das Verhältnis zweier zentraler Konzepte einer relationalen Betrachtung des Sozialen. In Fuhse, J., & Krenn, K. (Eds.), Netzwerke in gesellschaftlichen Feldern (pp. 63–79). Springer Fachmedien Wiesbaden, Wiesbaden. https://doi.org/10.1007/978-3-658-22215-4_3.

Scott, J. (2012). *Social network analysis*. New York: SAGE.

Shibata, N., Kajikawa, Y., & Matsushima, K. (2007). Topological analysis of citation networks to discover the future core articles. *Journal of the American Society for Information Science and Technology*, *58*(6), 872–882. https://doi.org/10.1002/asi.20529.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B. J. P., & Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web—WWW '15 Companion* (pp. 243–246). ACM Press. https://doi.org/10.1145/2740908.2742839.

Sjögårde, P., & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics*, *12*(1), 133–152. https://doi.org/10.1016/j.joi.2017.12.006.

Small, H. G. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269. https://doi.org/10.1002/asi.4630240406.

Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science*, *8*(3), 327–340. https://doi.org/10.1177/030631277800800305.

Stier, S., Bleier, A., Bonart, M., Mörsheim, F., Bohlouli, M., Nizhegorodov, M., Posch, L., Maier, J., Rothmund, T., & Staab, S. (2018). Systematically monitoring social media: The case of the German federal election 2017. *GESIS Papers*, *2018/04*. https://doi.org/10.21241/ssoar.56149.

Swidler, A. (1986). Culture in action: Symbols and strategies. *American Sociological Review*, *51*(2), 273–286. https://doi.org/10.2307/2095521.

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. https://doi.org/10.1002/asi.22748.

West, G. (2017). *Scale: The universal laws of growth, innovation, sustainability, and the pace of life in organisms, cities, economies, and companies*. New York, NY: Penguin Press.

White, H. C. (2008). *Identity and control: How social formations emerge*. Princeton, NJ: Princeton University Press.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099.

Zipf, G. K. (2012 [1949]). Human behaviour and the principle of least effort: An introduction to human ecology. Mansfield Centre, CT: Martino.

Zitt, M. (2015). Meso-level retrieval: IR-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation. *Scientometrics*, *102*(3), 2223–2245. https://doi.org/10.1007/s11192-014-1482-5.

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing & Management*, *42*(6), 1513–1531. https://doi.org/10.1016/j.ipm.2006.03.016.

Zuccala, A. (2006). Modeling the invisible college. *Journal of the American Society for Information Science and Technology*, *57*(2), 152–168. https://doi.org/10.1002/asi.20256.